# Capstone Project-3

## Mobile Price Range Prediction

Team Members

## Team Members

1 Arvind  Kale

2 Priyanka Jain

3 Mohd Sakib Quraishi

4.Nitish kumar

# Contents :

# Problem Statement

**The problem statement is to predict the price range of mobile phones based on the features available (price range indicating how high the price is). Here is the description of target classes:**

- **0 - Low cost Phones**
- **1 - Medium cost phones**
- **2 - High cost phones**
- **3 - Very High cost phones**

**This will basically help companies to estimate price of mobiles to give tough competition to other mobile manufacturer.**

**Also, it will be useful for consumers to verify that they are paying best price for a mobile.**

# Attribute Information

- Battery_power - Total energy a battery can store in one time measured in mAh
- Blue - Has bluetooth or not
- Clock_speed - speed at which microprocessor executes instructions
- Dual_sim - Has dual sim support or not
- Fc - Front Camera megapixels
- Four_g - Has 4G or not
- Int_memory - Internal Memory in Gigabytes
- M_dep - Mobile Depth in cm
- Mobile_wt - Weight of mobile phone
- N_cores - Number of cores of processor
- Pc - Primary Camera megapixels
- Px_height - Pixel Resolution Height
- Px_width - Pixel Resolution Width
- Ram - Random Access Memory in MegaBytes
- Sc_h - Screen Height of mobile in cm
- Sc_w - Screen Width of mobile in cm
- Talk_time - longest time that a single battery charge will last
- Three_g - Has 3G or not
- Touch_screen - Has touch screen or not
- Wifi - Has wifi or not ● Price_range - This is the target variable with value of 0(low cost), 1(medium cost),2(high cost) and 3(very high cost).

# Data Inspection:

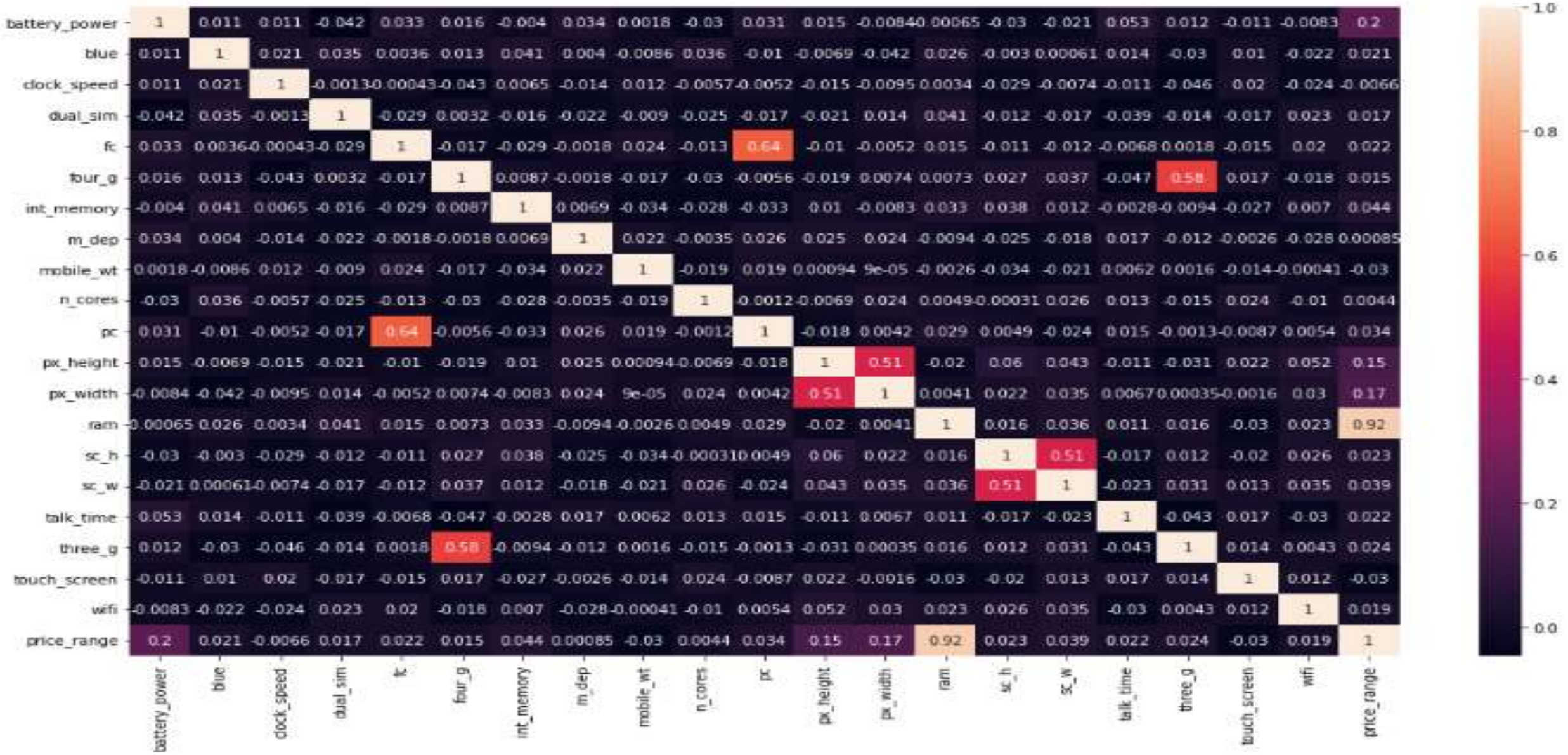
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   battery_power   2000 non-null   int64  
 1   blue            2000 non-null   int64  
 2   clock_speed     2000 non-null   float64
 3   dual_sim        2000 non-null   int64  
 4   fc              2000 non-null   int64  
 5   four_g          2000 non-null   int64  
 6   int_memory      2000 non-null   int64  
 7   m_dep           2000 non-null   float64
 8   mobile_wt       2000 non-null   int64  
 9   n_cores         2000 non-null   int64  
 10  pc              2000 non-null   int64  
 11  px_height       2000 non-null   int64  
 12  px_width        2000 non-null   int64  
 13  ram             2000 non-null   int64  
 14  sc_h            2000 non-null   int64  
 15  sc_w            2000 non-null   int64  
 16  talk_time       2000 non-null   int64  
 17  three_g         2000 non-null   int64  
 18  touch_screen    2000 non-null   int64  
 19  wifi            2000 non-null   int64  
 20  price_range     2000 non-null   int64  
dtypes: float64(2), int64(19)
memory usage: 328.2 KB
```

```
[6]  df.isnull().sum()
```

```
battery_power    0
blue             0
clock_speed      0
dual_sim         0
fc               0
four_g           0
int_memory       0
m_dep            0
mobile_wt        0
n_cores          0
pc               0
px_height        0
px_width         0
ram              0
sc_h             0
sc_w             0
talk_time        0
three_g          0
touch_screen     0
wifi             0
price_range      0
dtype: int64
```

# Correlation Heatmap



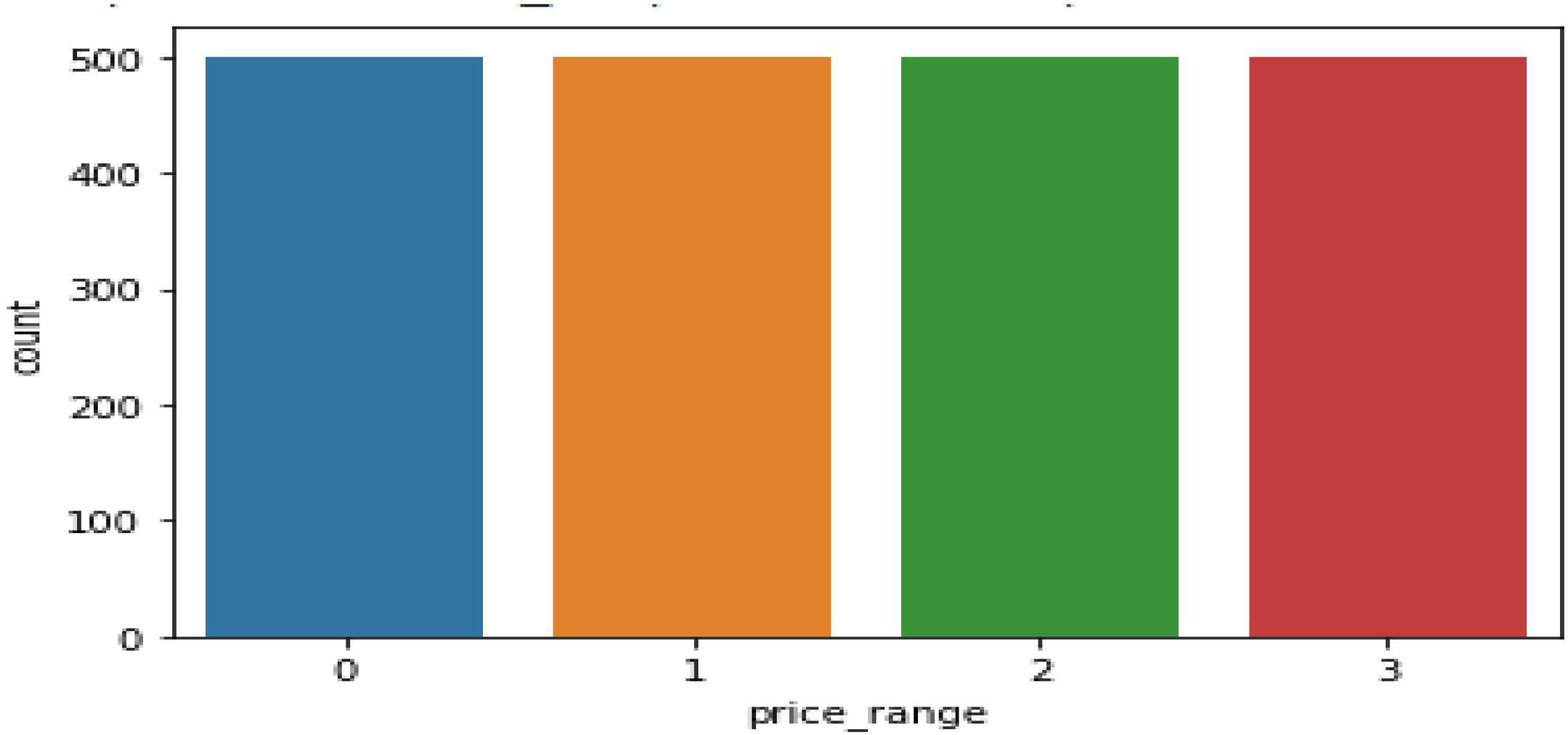

**There are no two column which are strongly related to each other but we will do some feature engineering to reduce the number of columns**

# Feature Engineering


- Generally the screen size of the phone is expressed in Inches.
- We have columns ‘sc_h’ and ‘sc_w’ out of which we have created a new feature ‘Screen_size’ which is diagonal length of the screen.

This will help to remove two column sc_h and sc_w.

# Exploratory Data Analysis

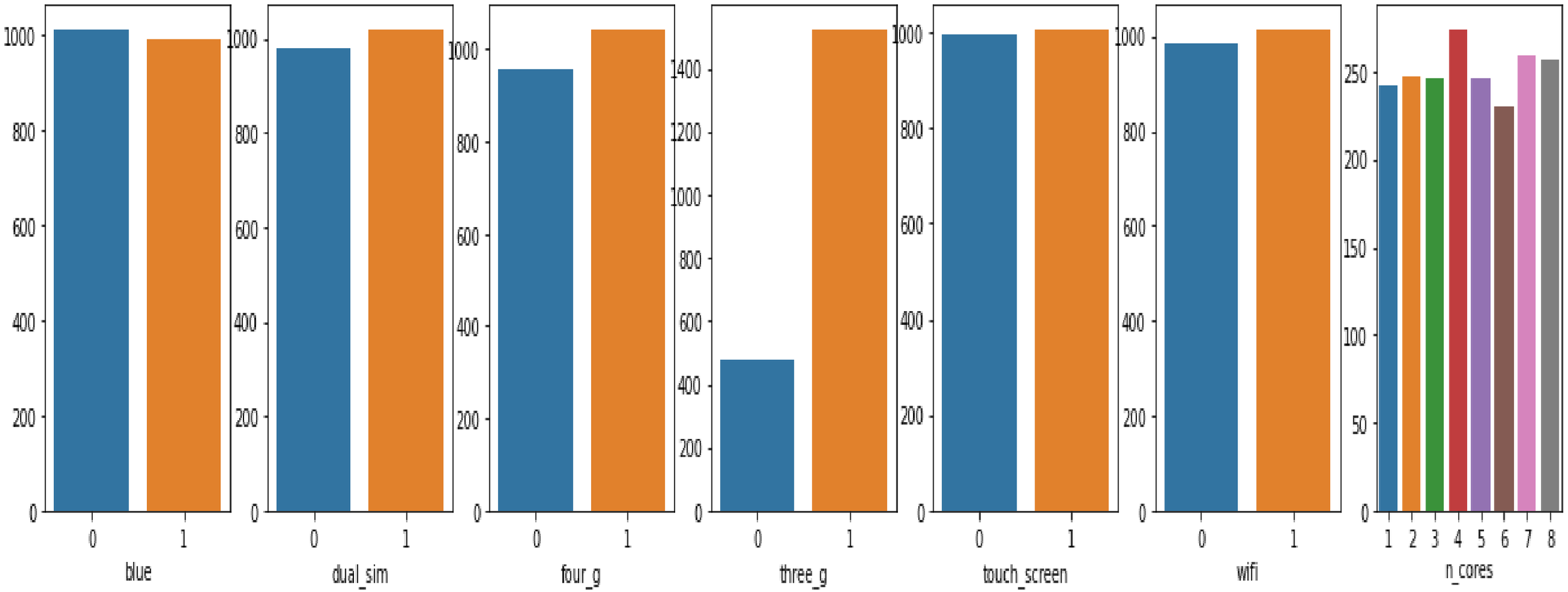



**We have perfectly balanced dataset with 500 observations for each class.**

# Exploratory Data Analysis on Price range and number of mobile phones


## Each price range have equal mobiles

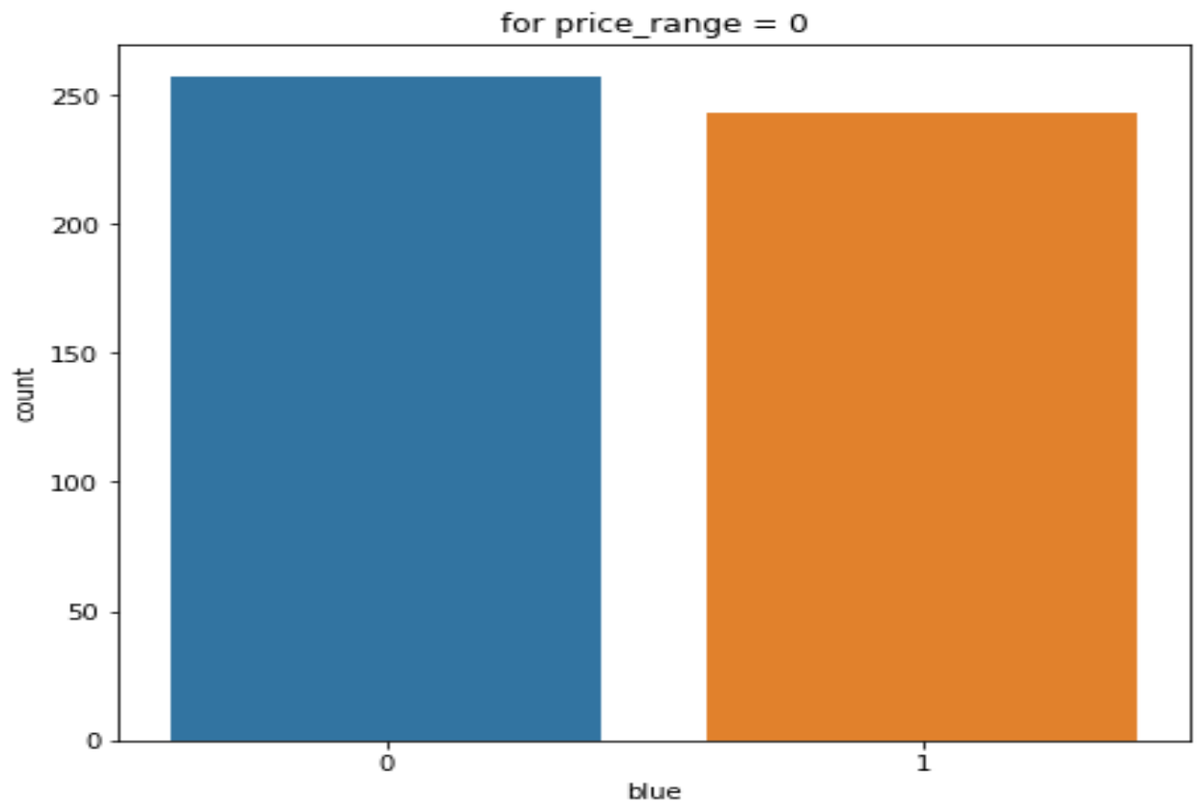
# Counting mobiles based on feature

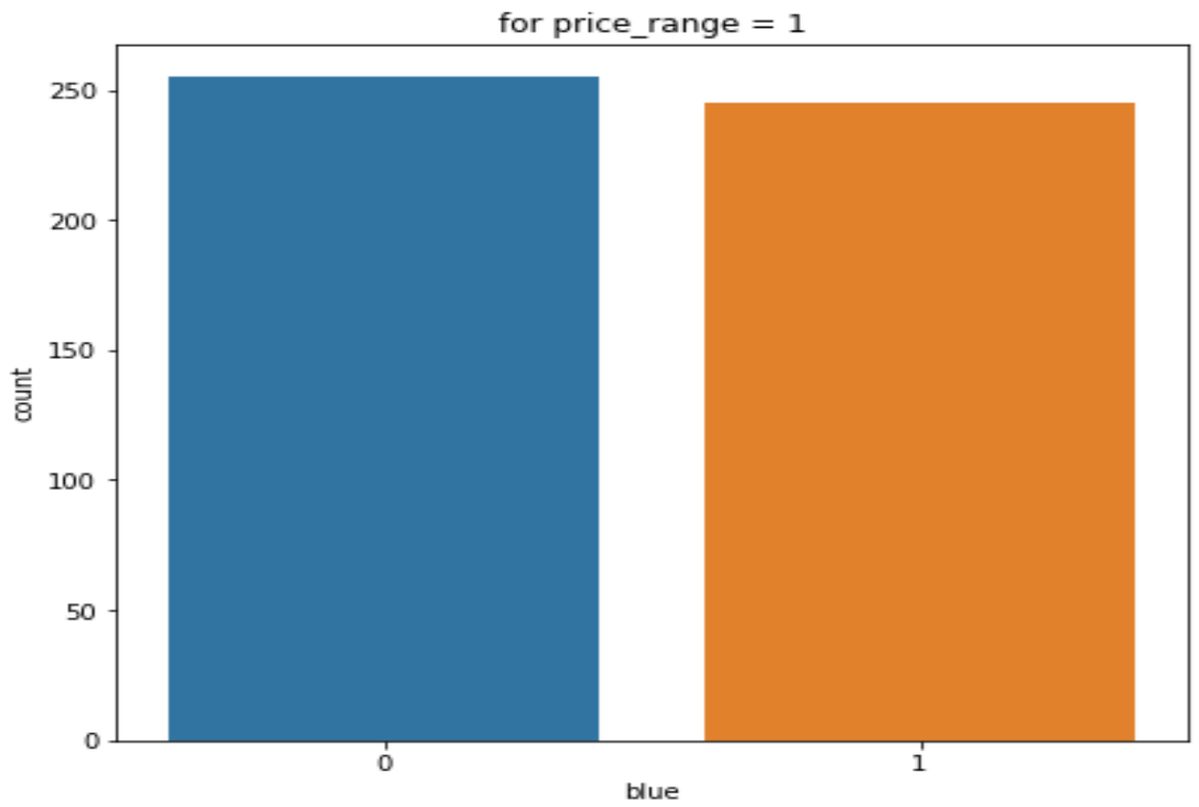


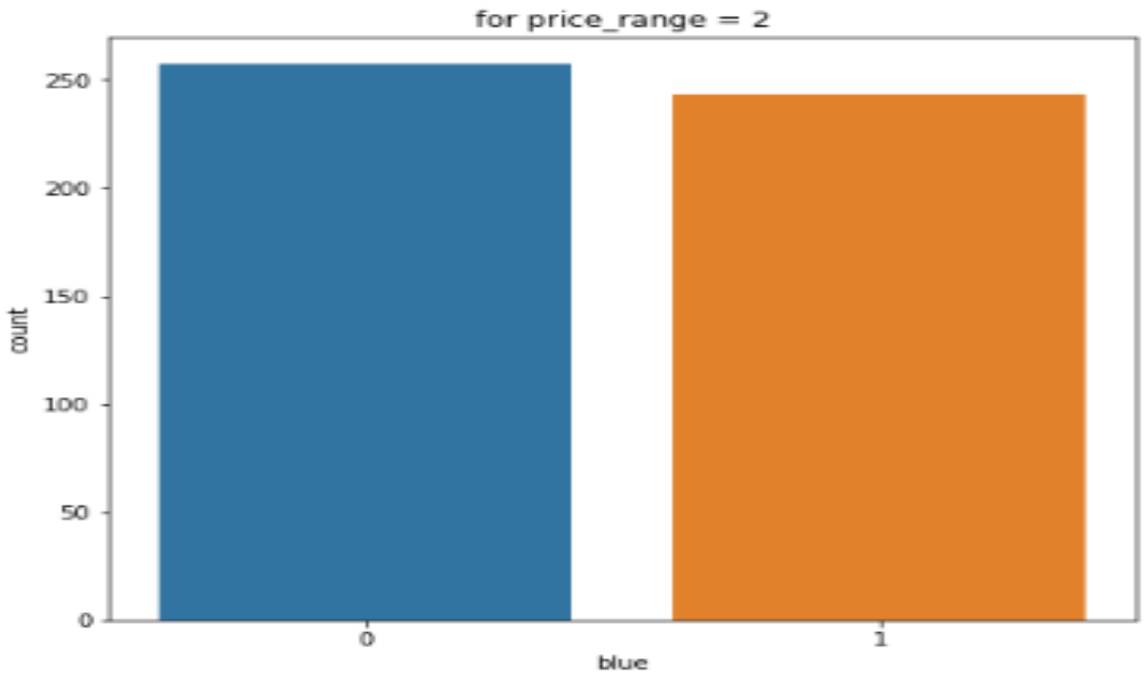

**. Below figure shows the categorical variables . The first one is whether the phone has bluetooth or not. It shows that half have bluetooth and half have not. Second variable is dual sim . It is also almost 50-50 . The only unbalanced variable here is 3g . Almost 500 variables don’t have 3g while another 1400 have.**

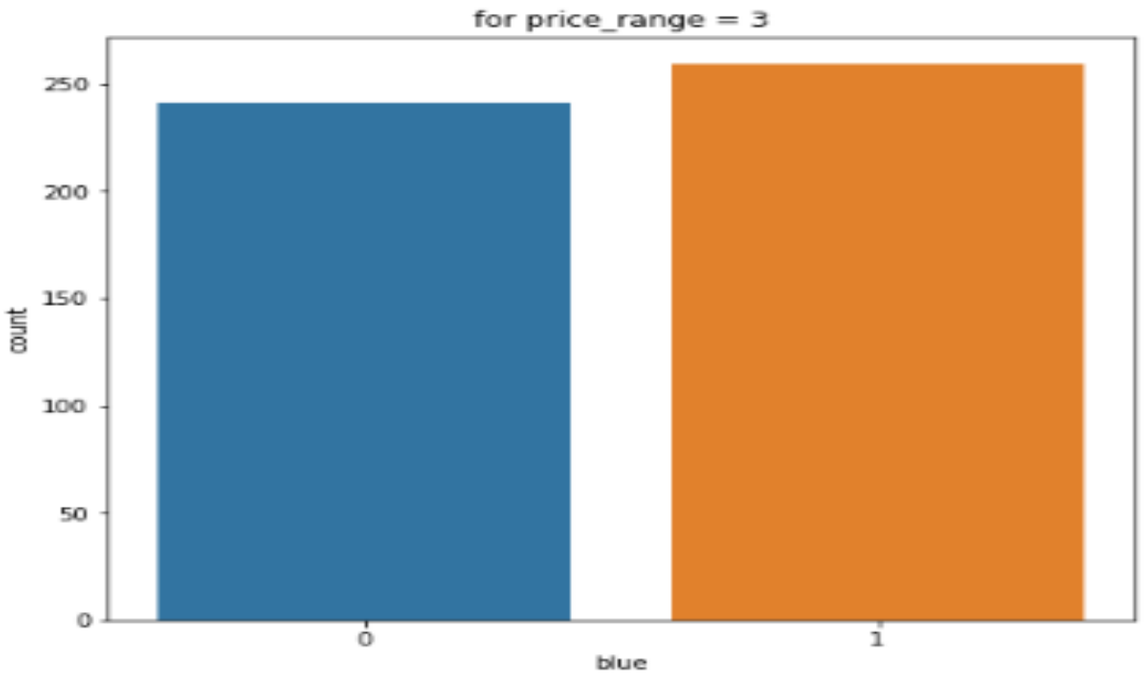
# Analysing that the number of phones with all features based on price range(0,1,2,3)

7

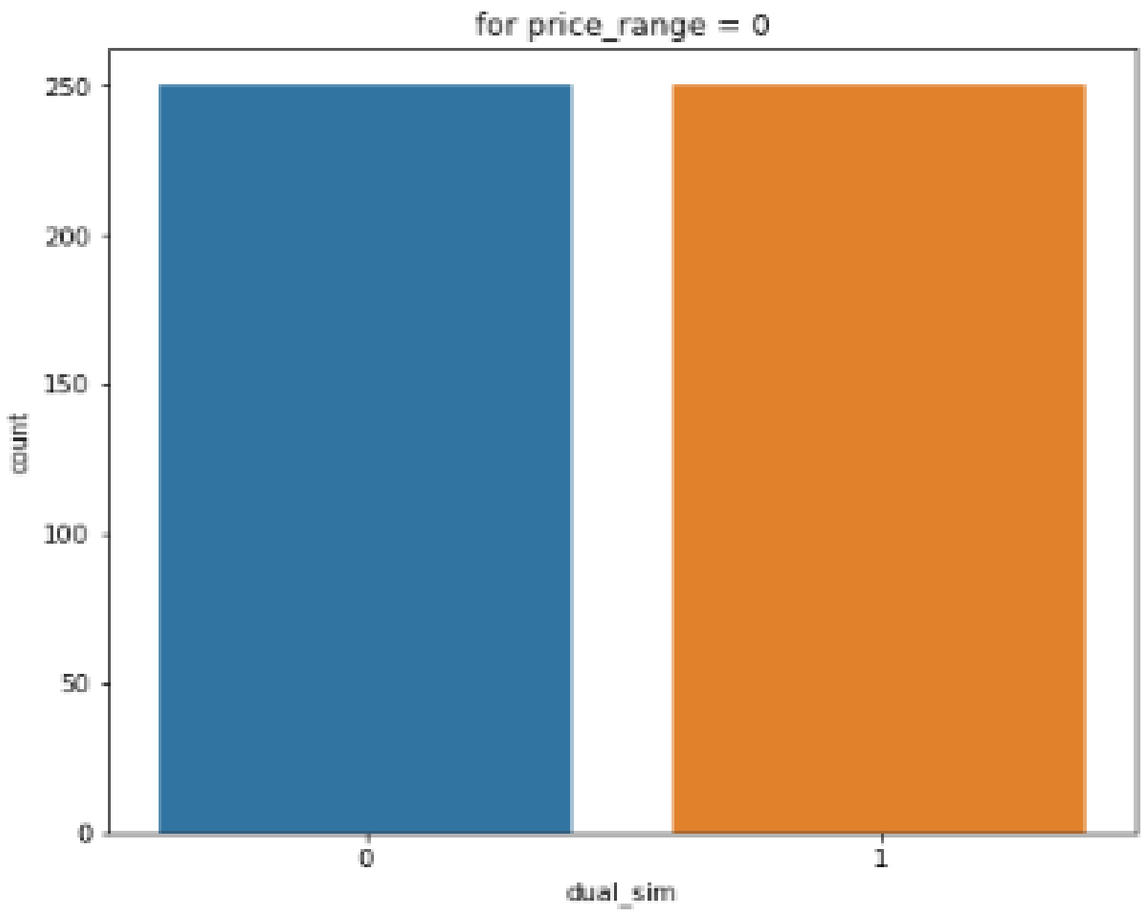


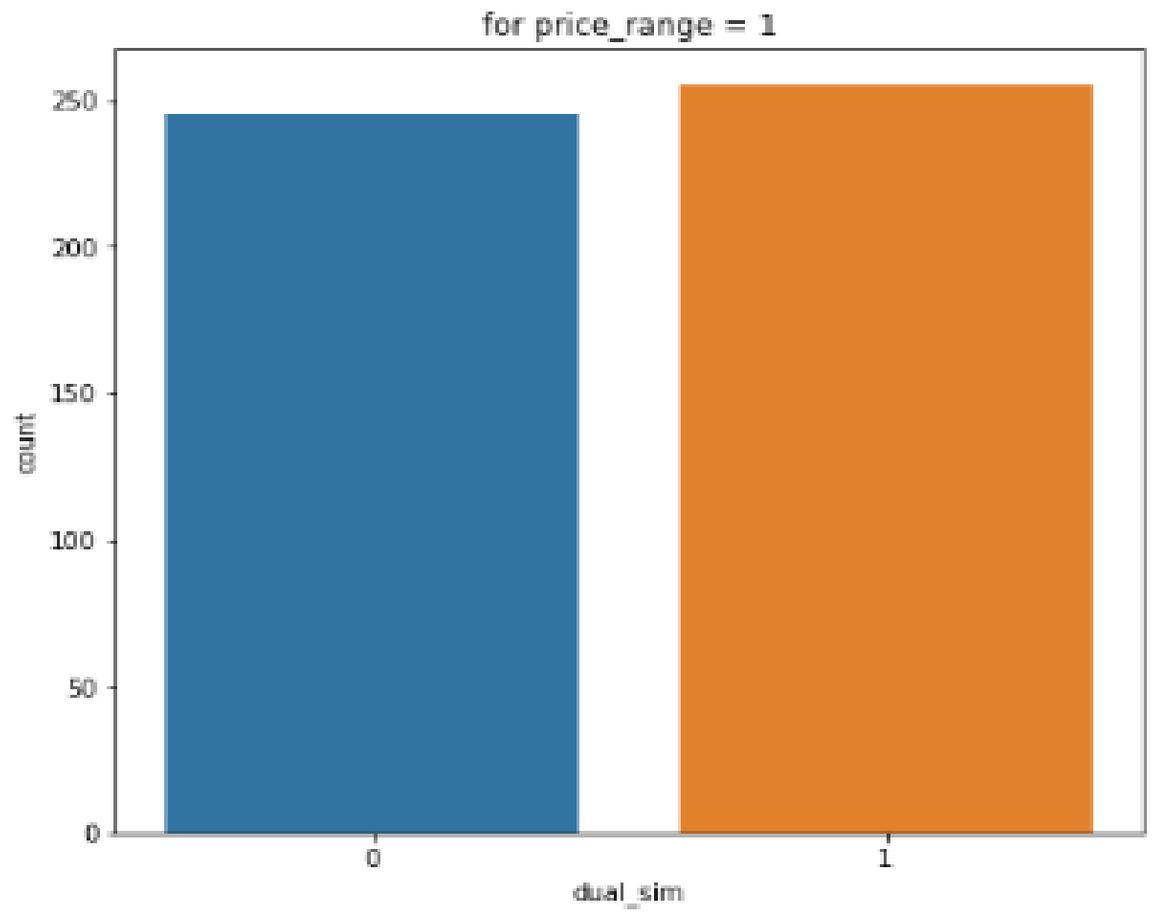



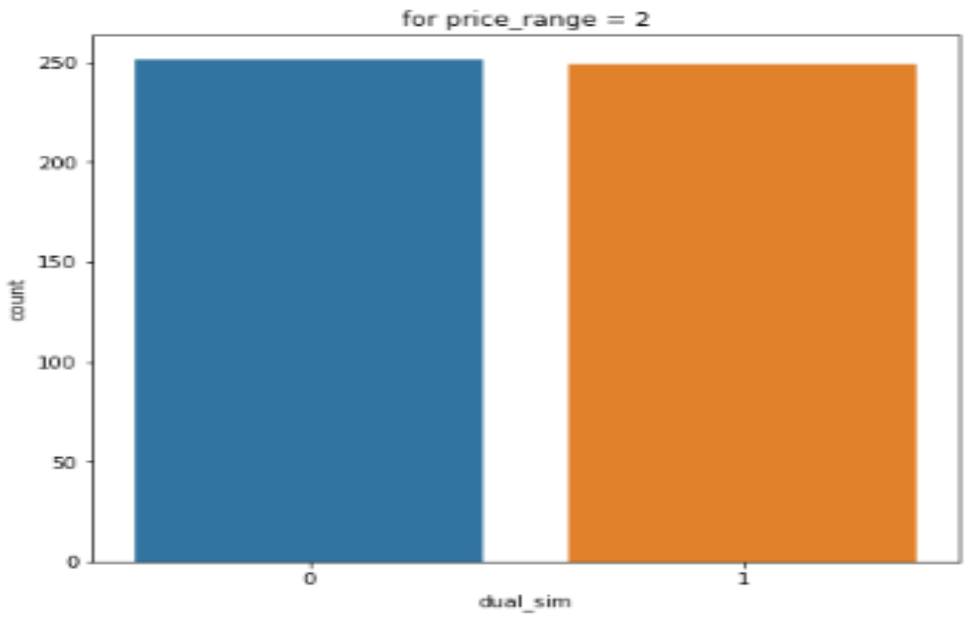



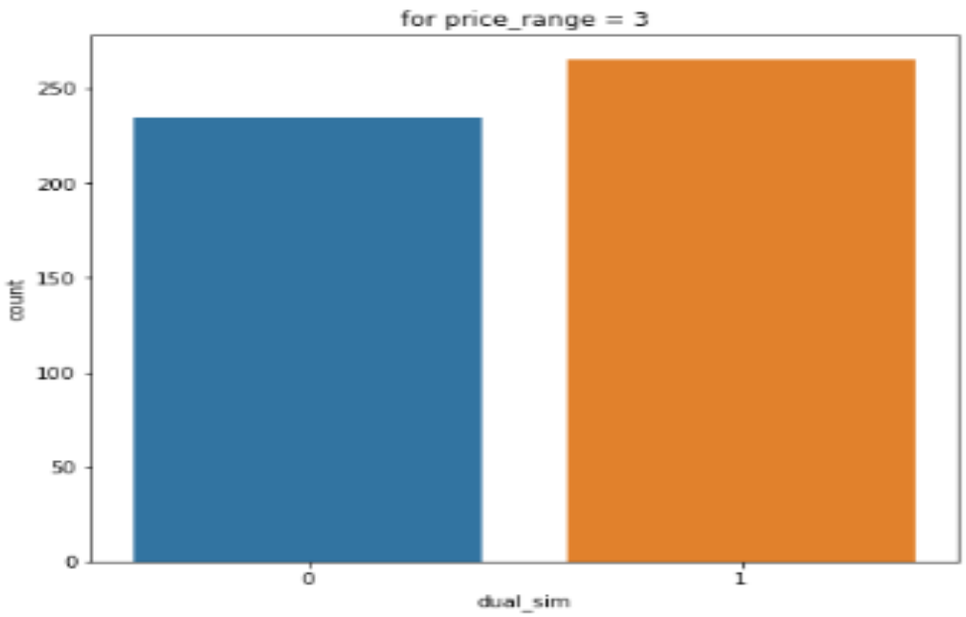



# Below figure shows that dual sim features are 50-50% available in all mobile price range.

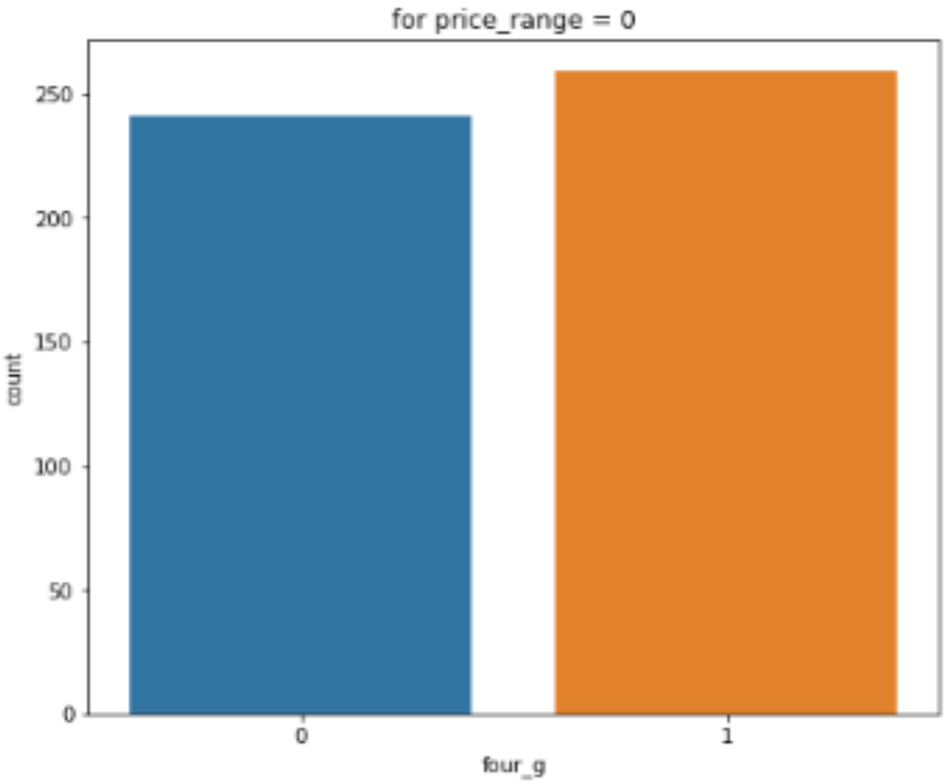


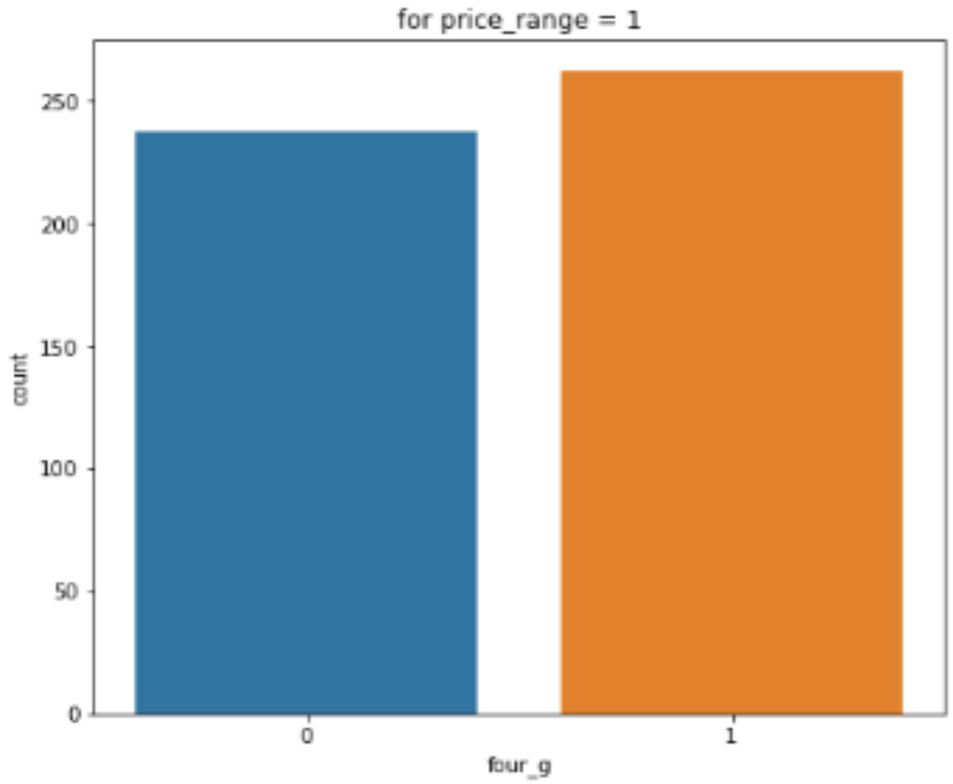



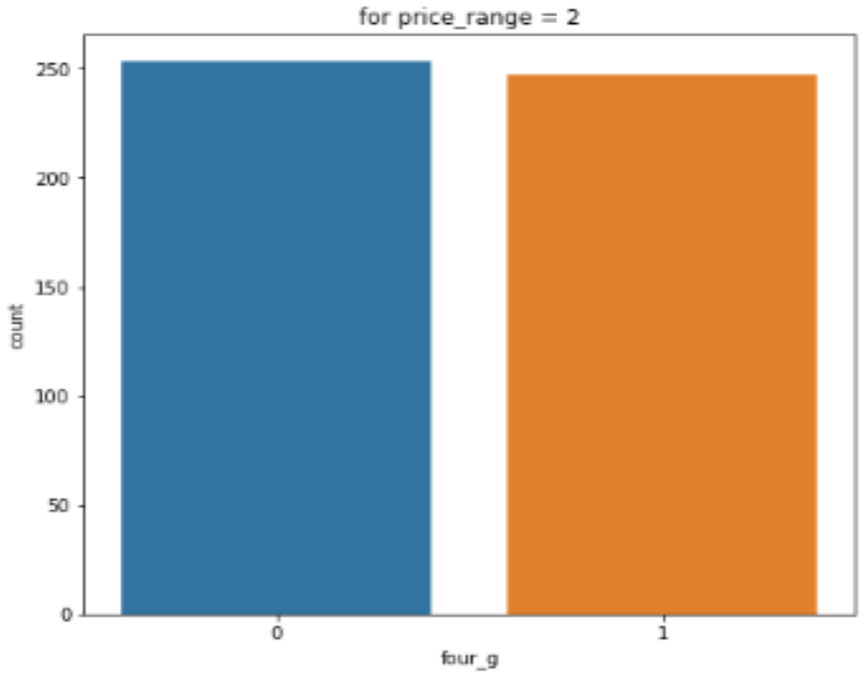



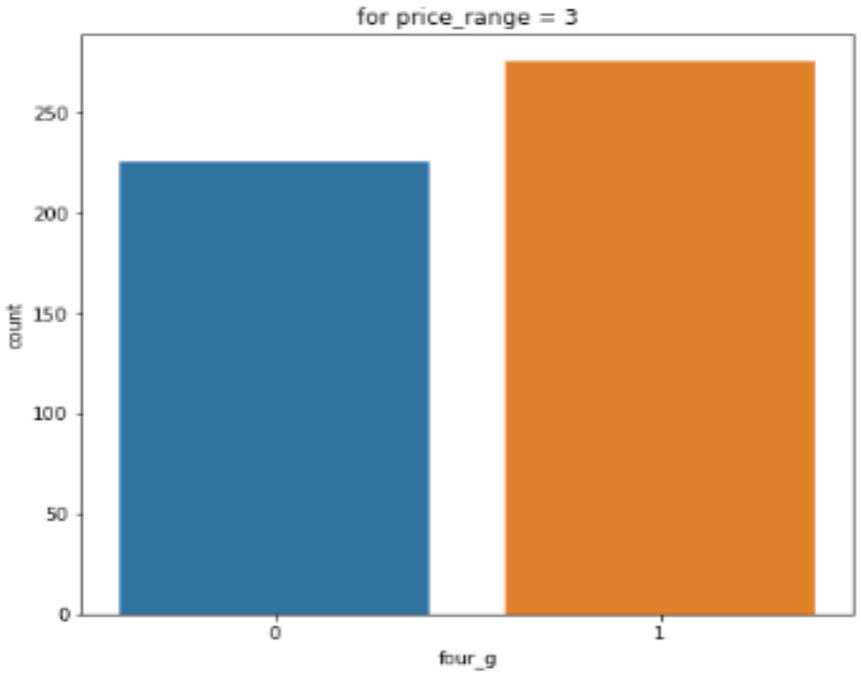



# Below graph shows that how many mobile have 4G in all price segment

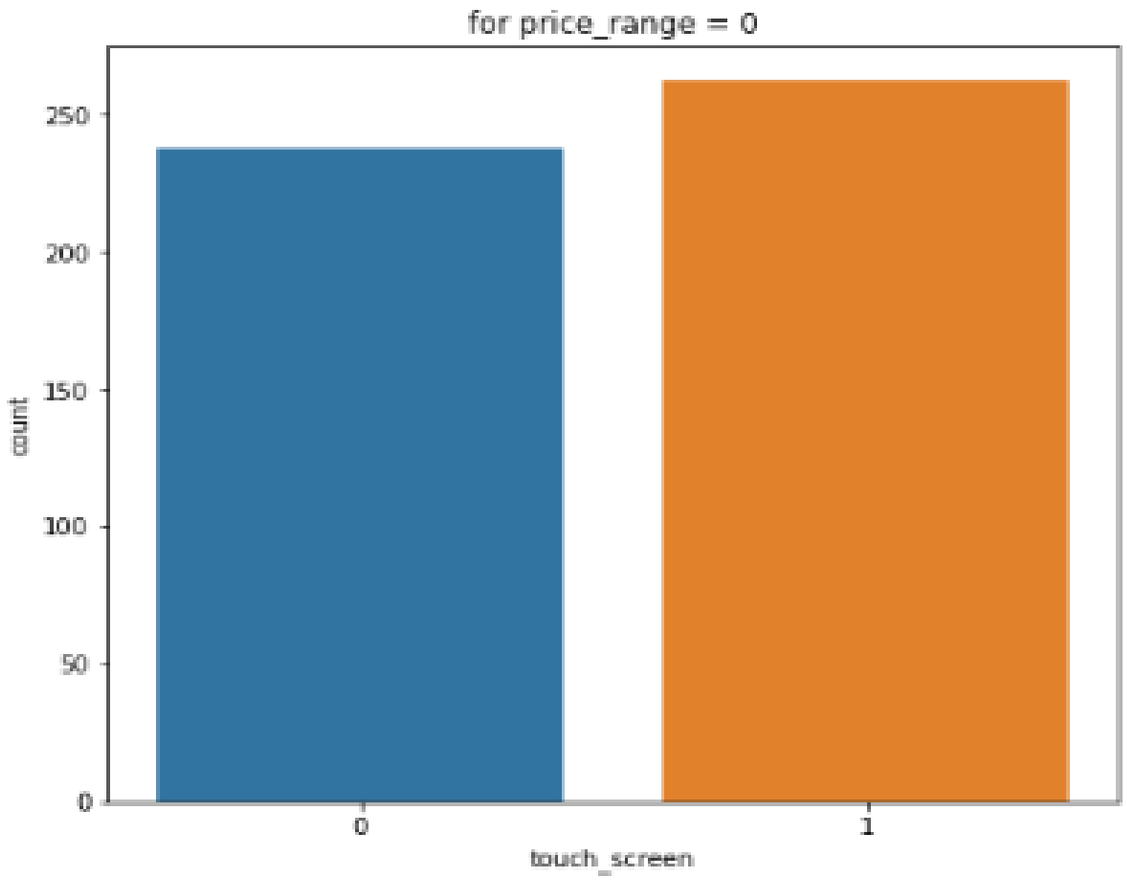


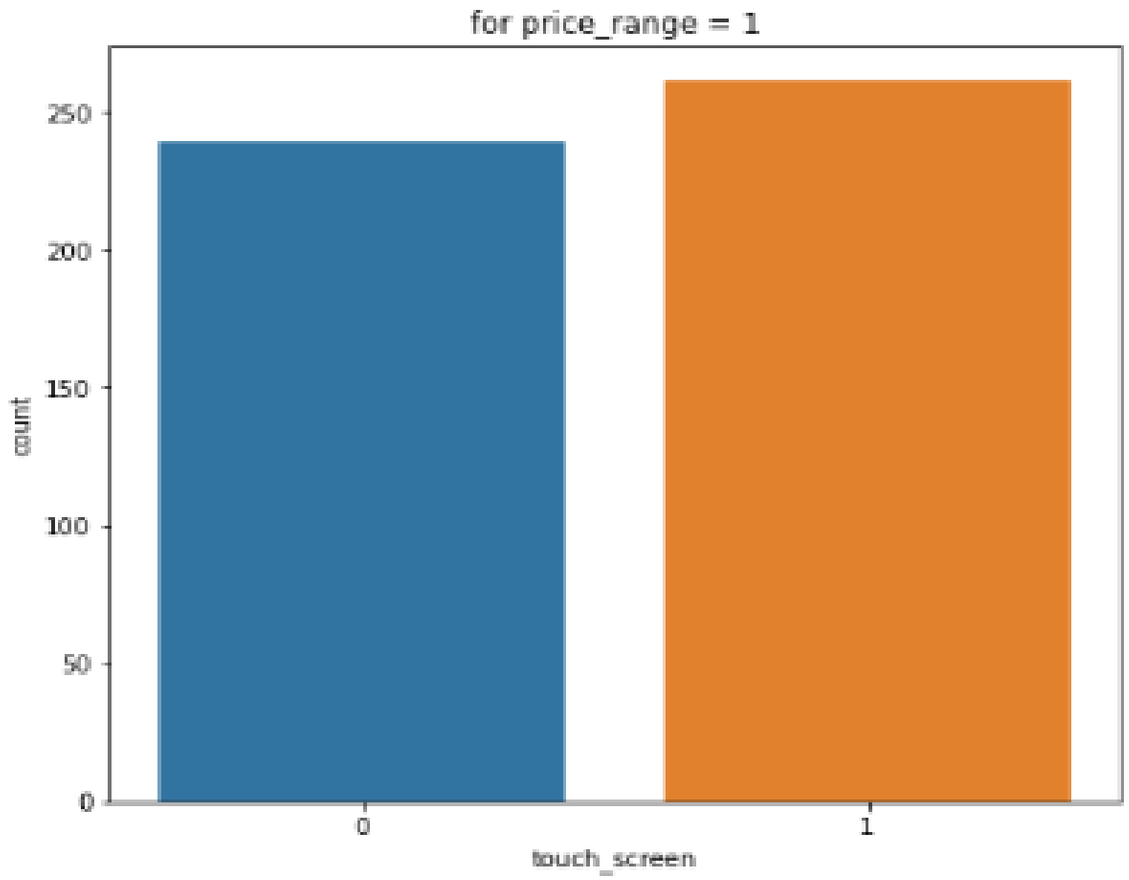



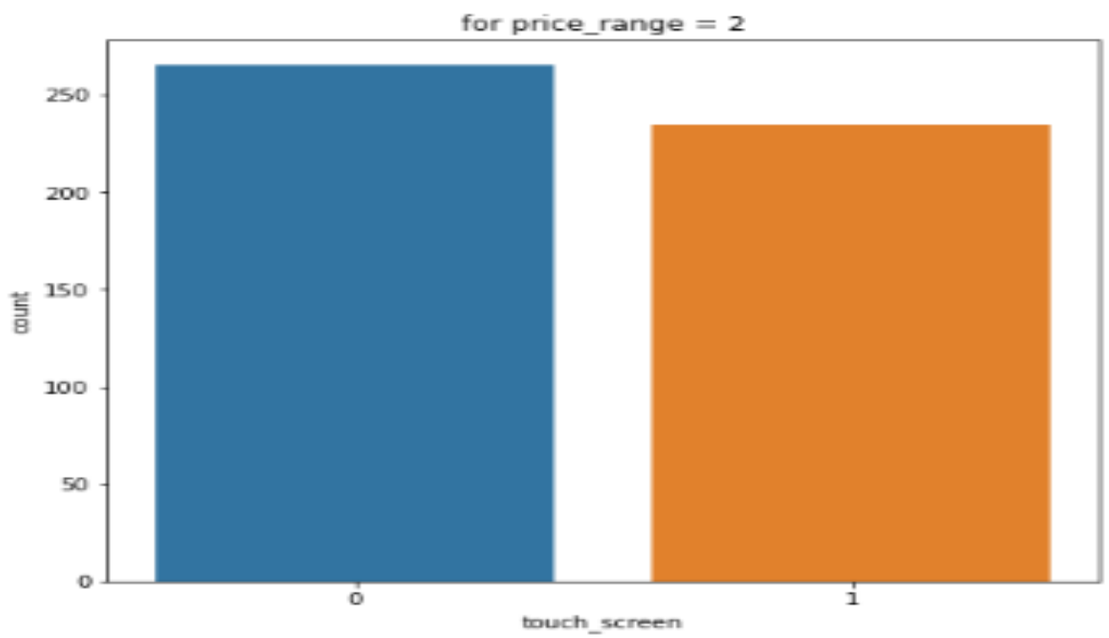



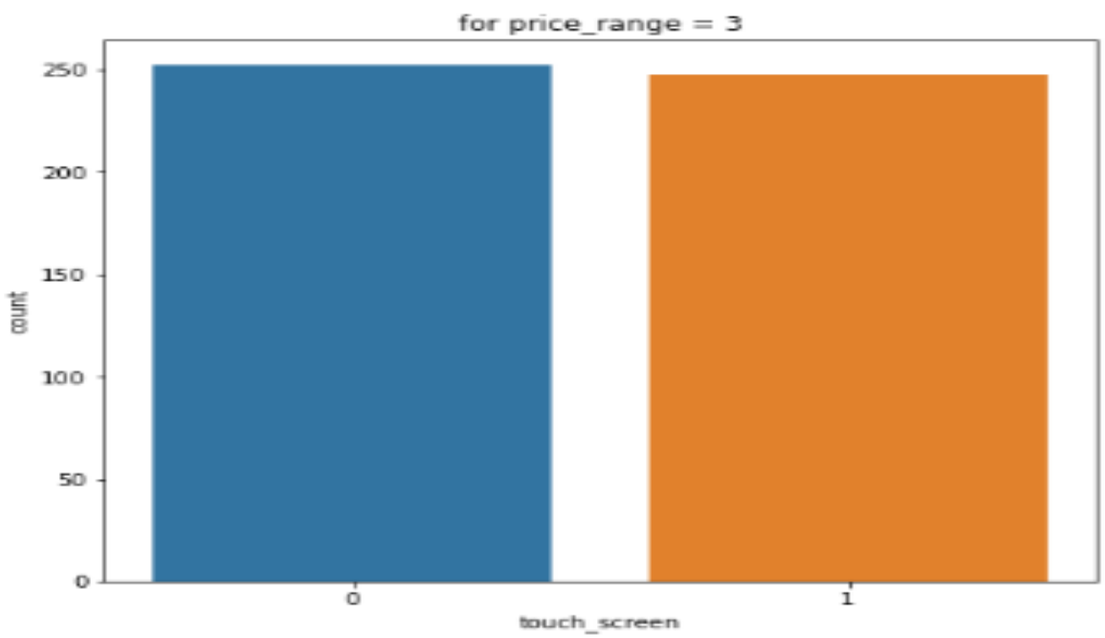



# Analysis of touch screen in all 4 price range of mobile

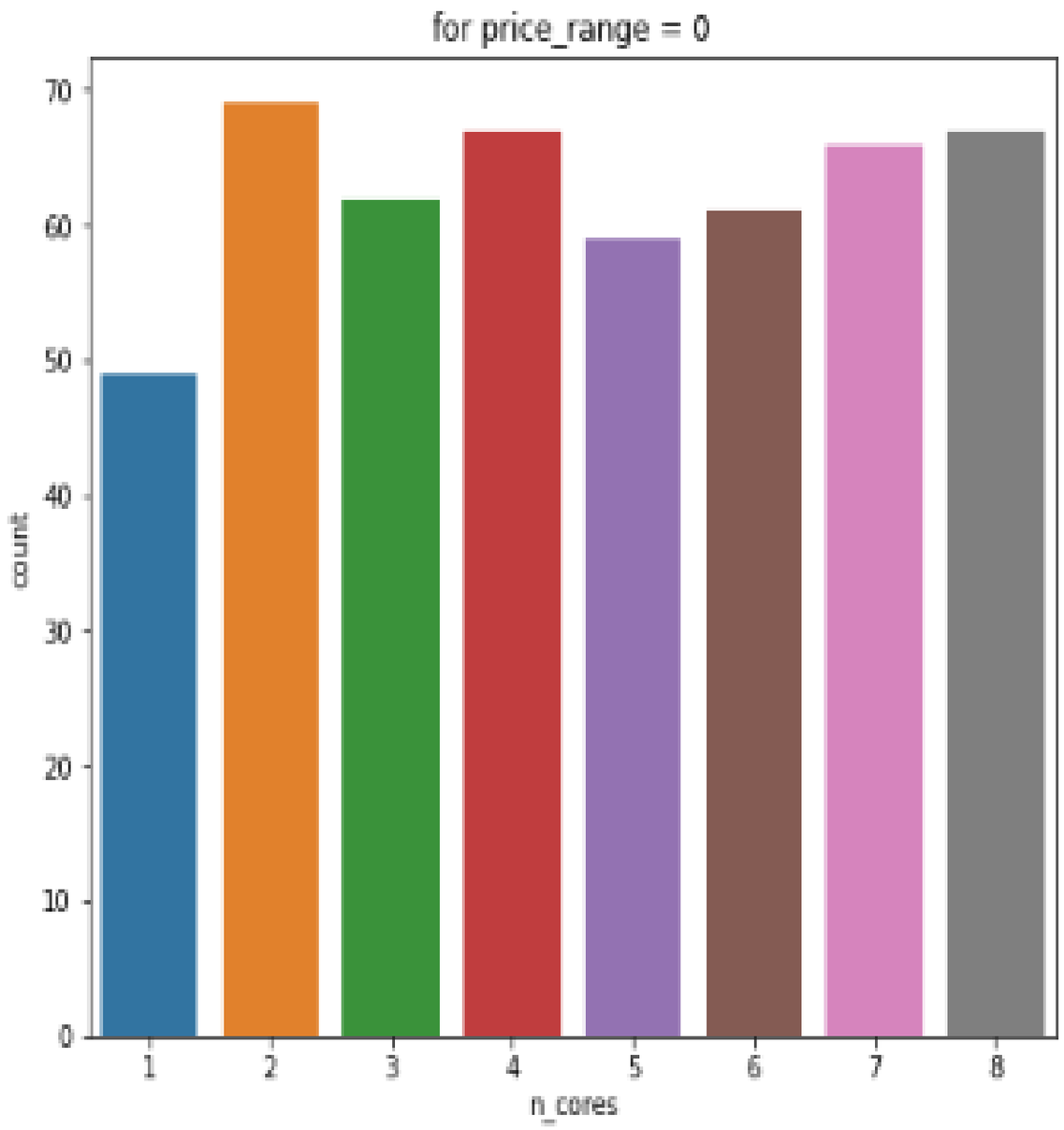


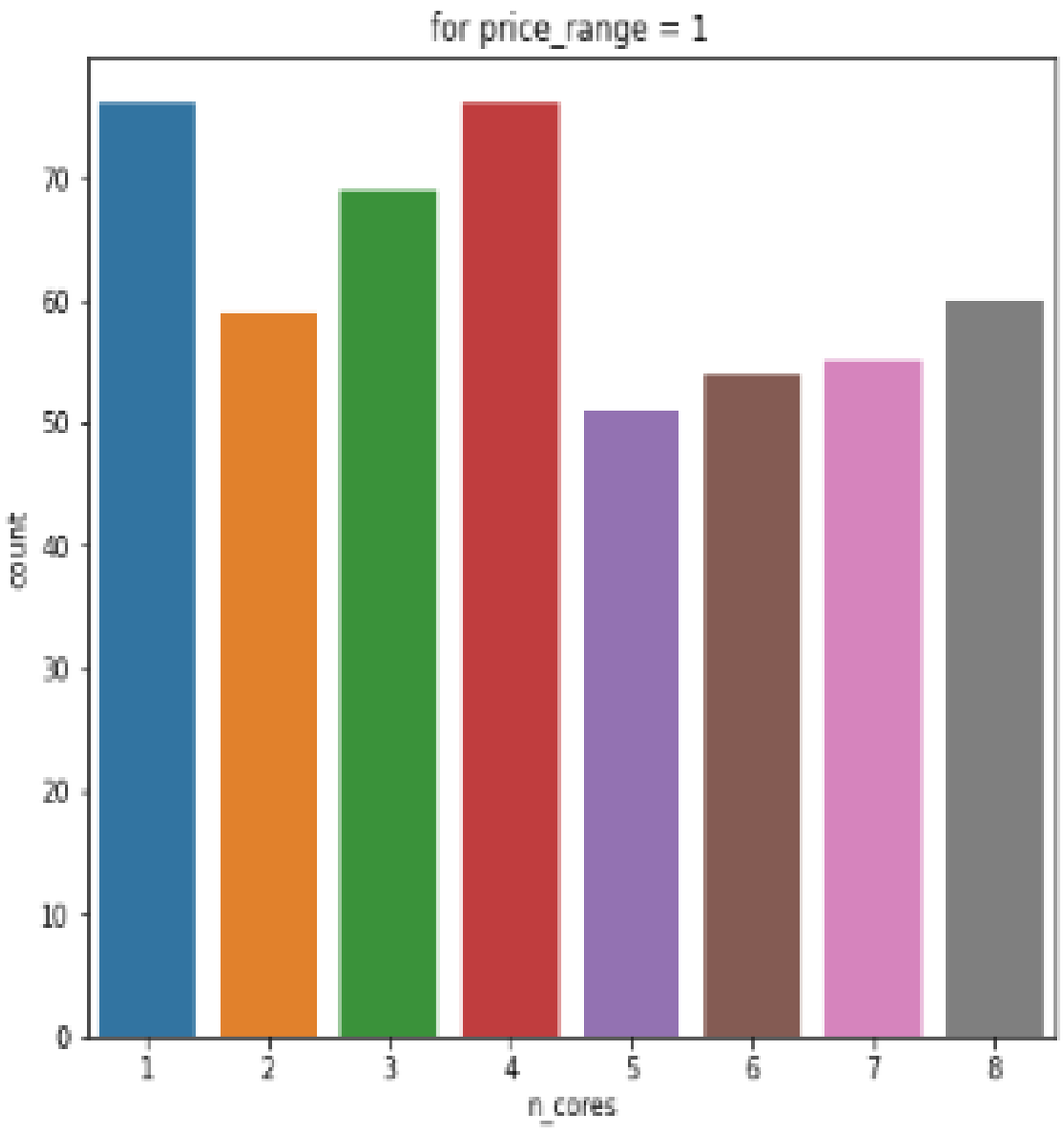



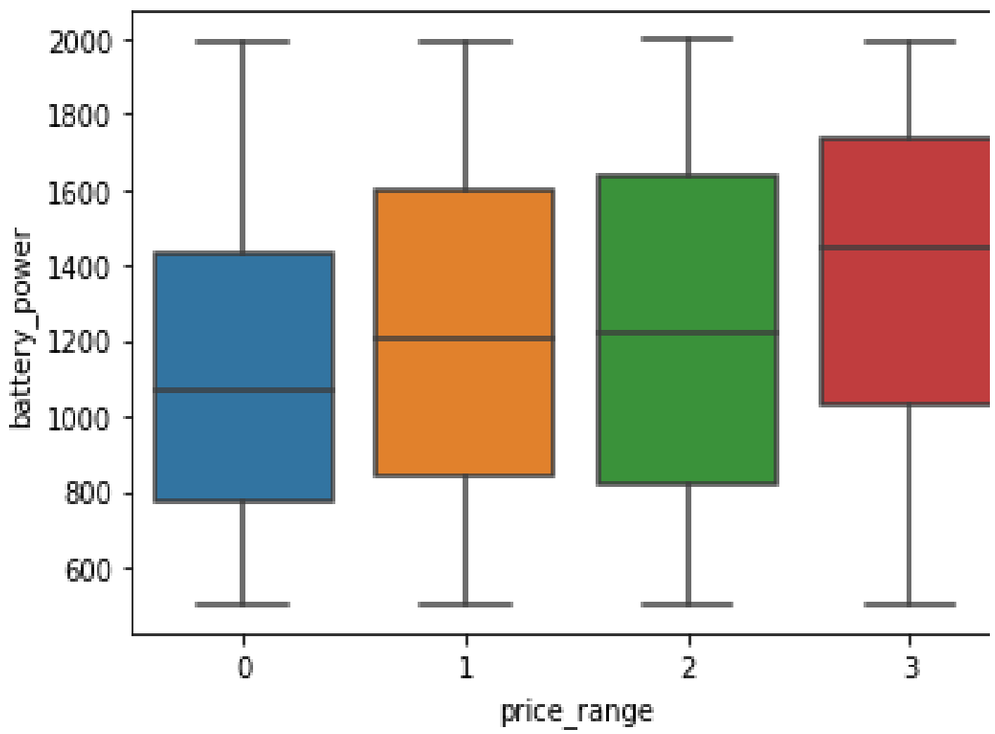



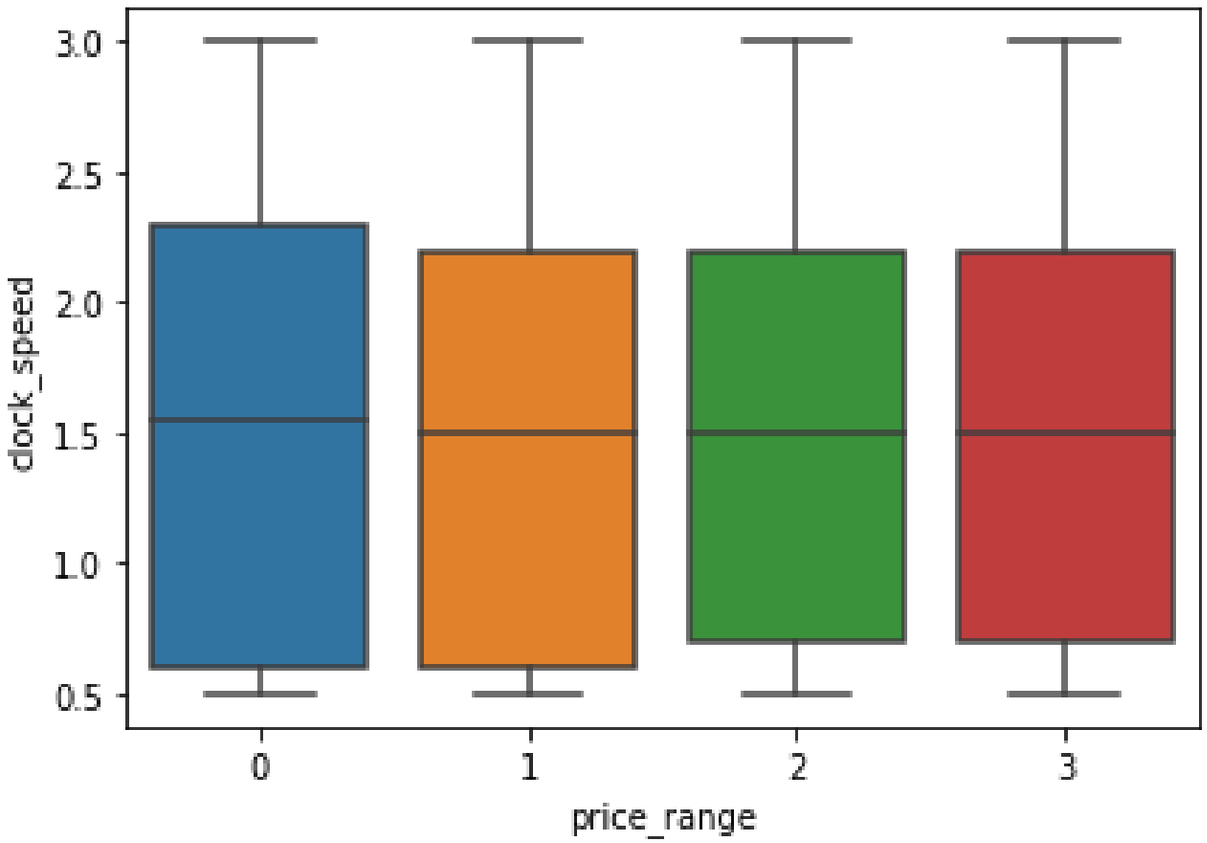
# 2nd and 4th cores features shows max count for 0 price range and 1st and 4th cores shows max count for 1 price range

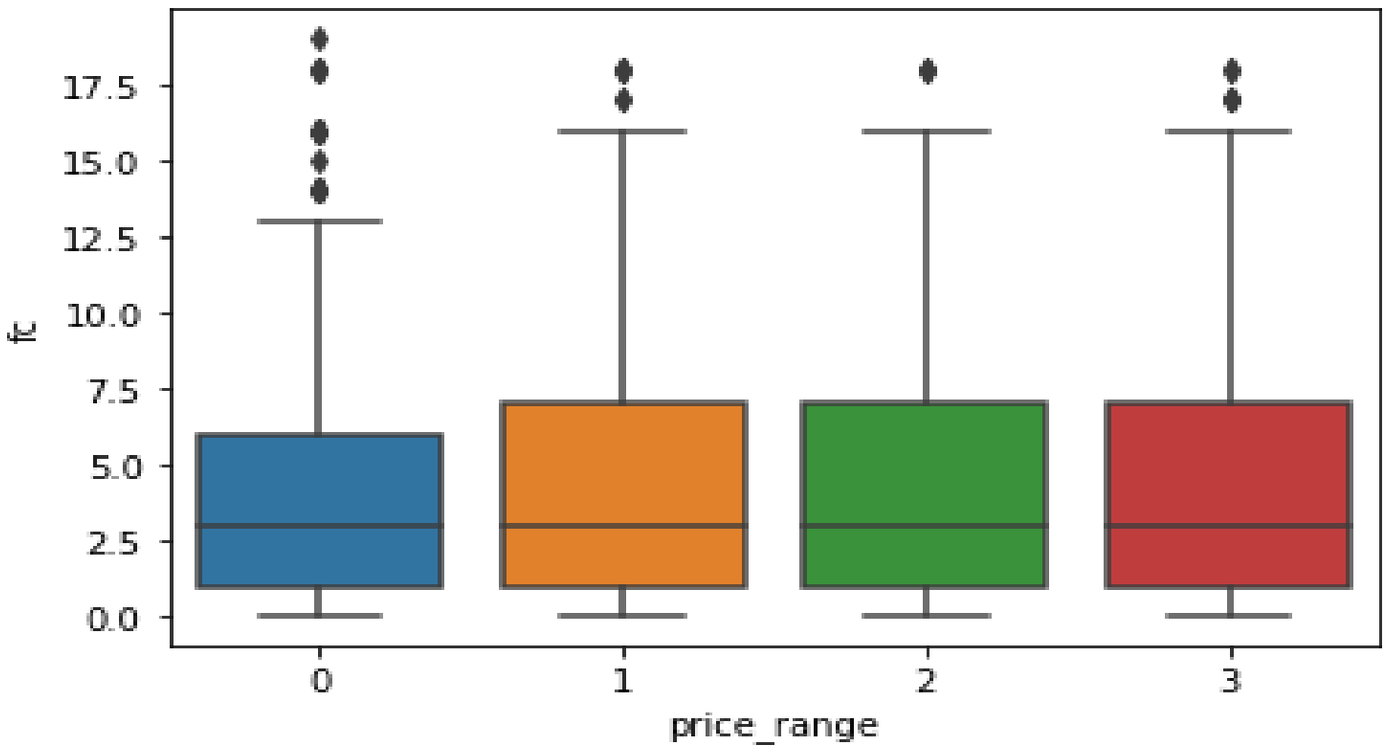


# Analysis based on price range for different feature



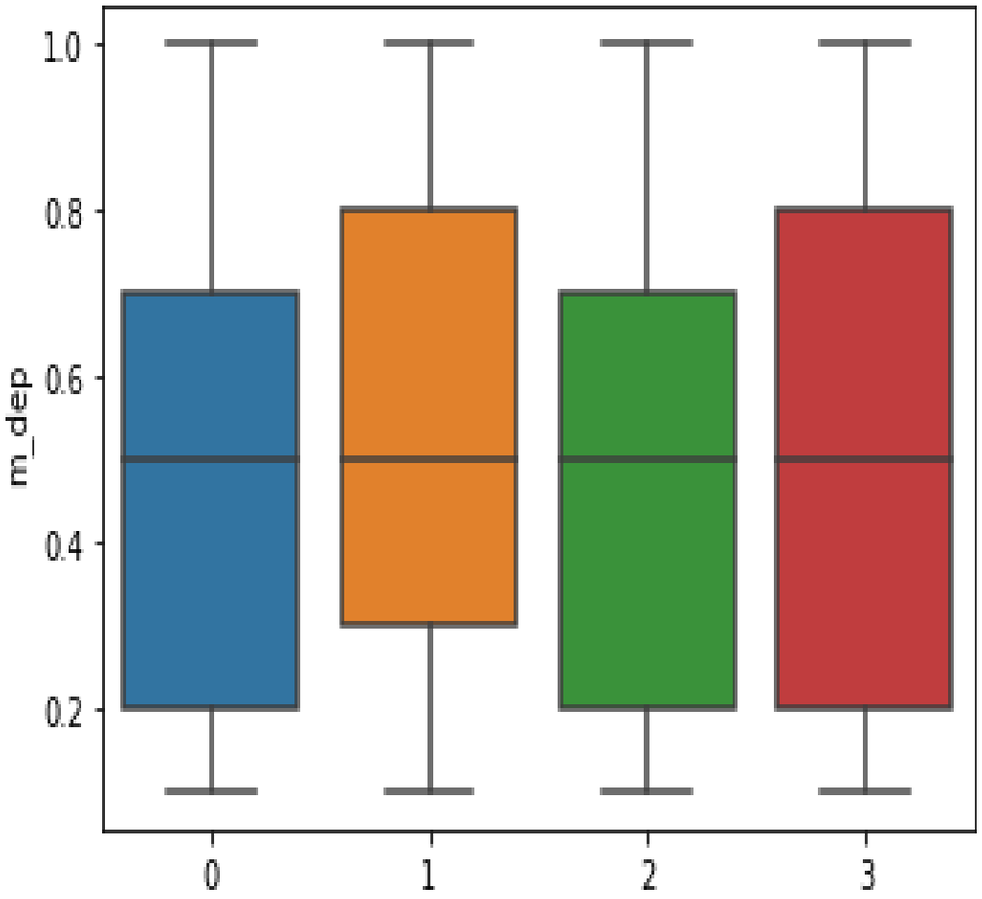



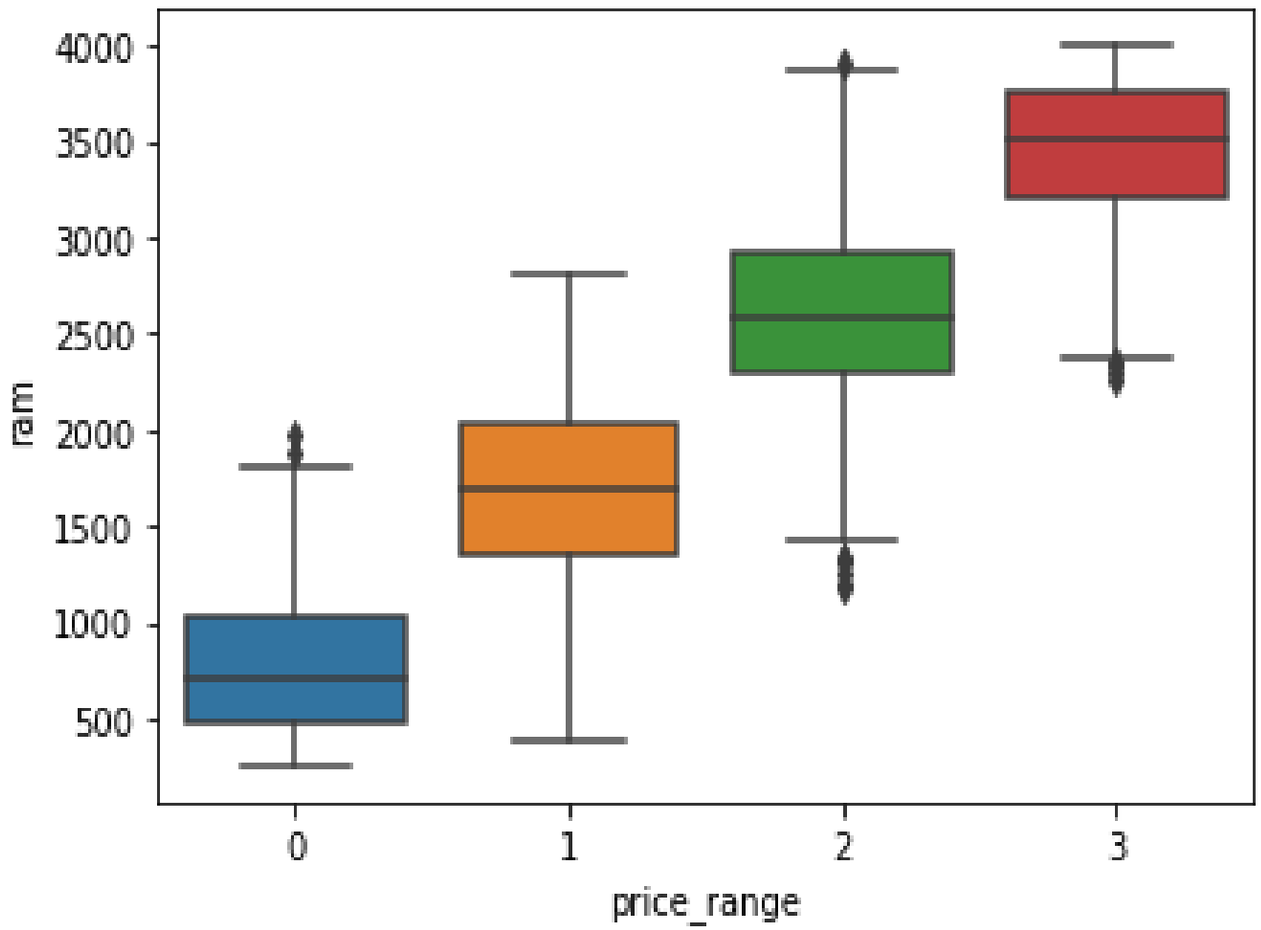



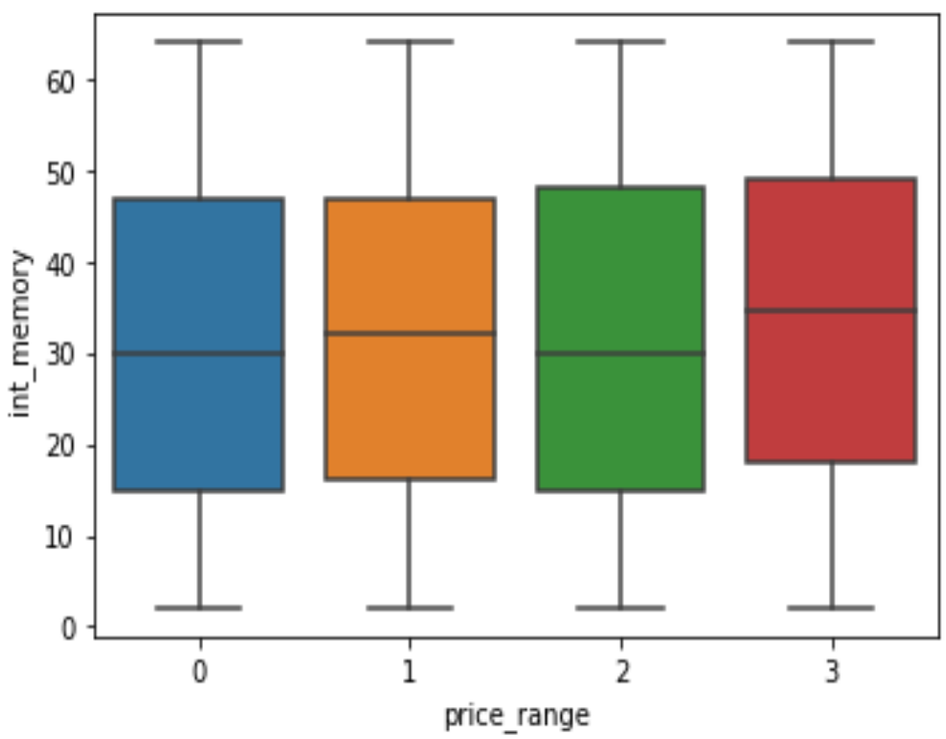

**Battery power show the most variation along the different price ranges.**

**Clock speed and fc show the almost 50% variation along the different price range.**

# Analysis of different feature for different price of mobile



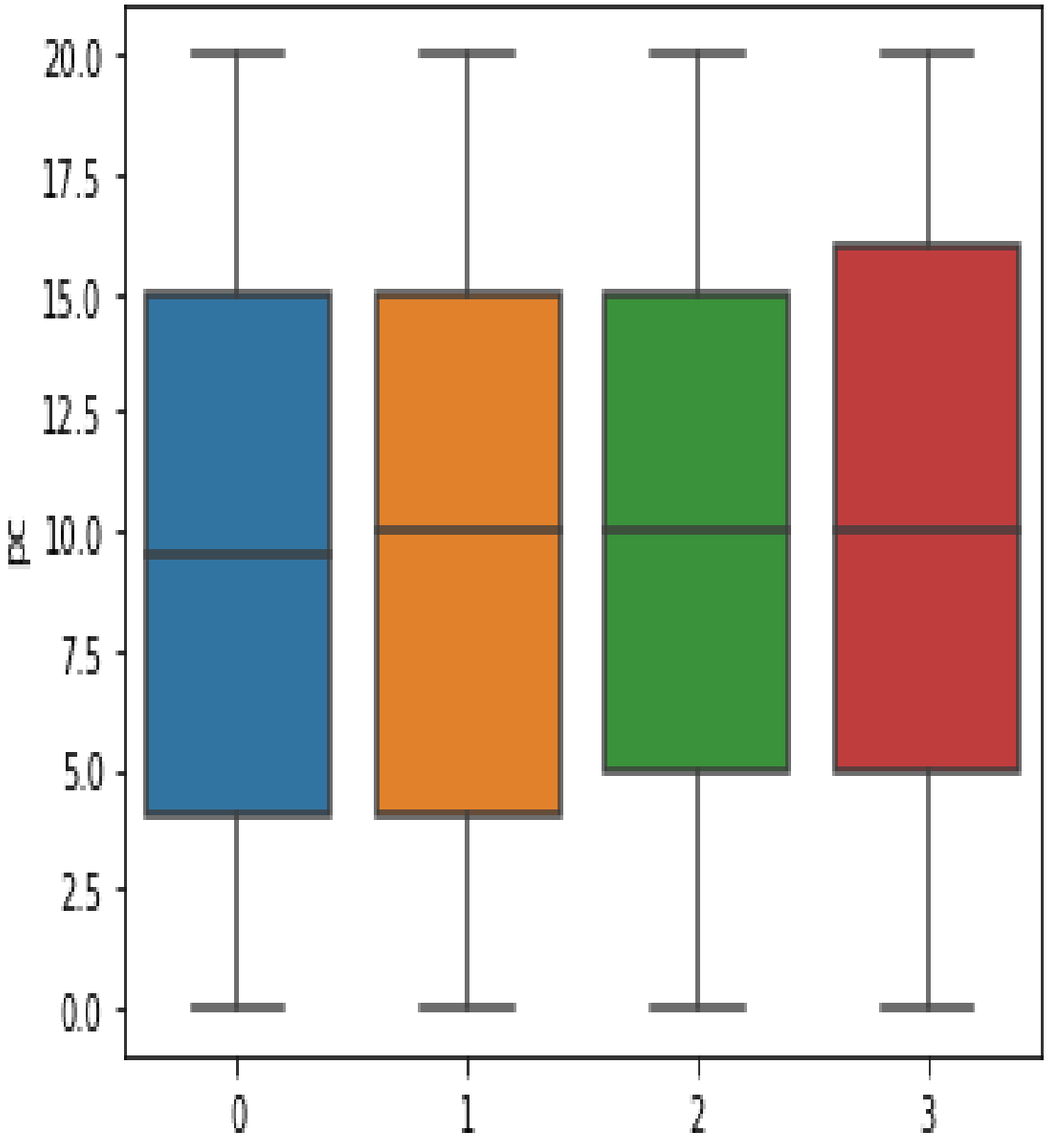



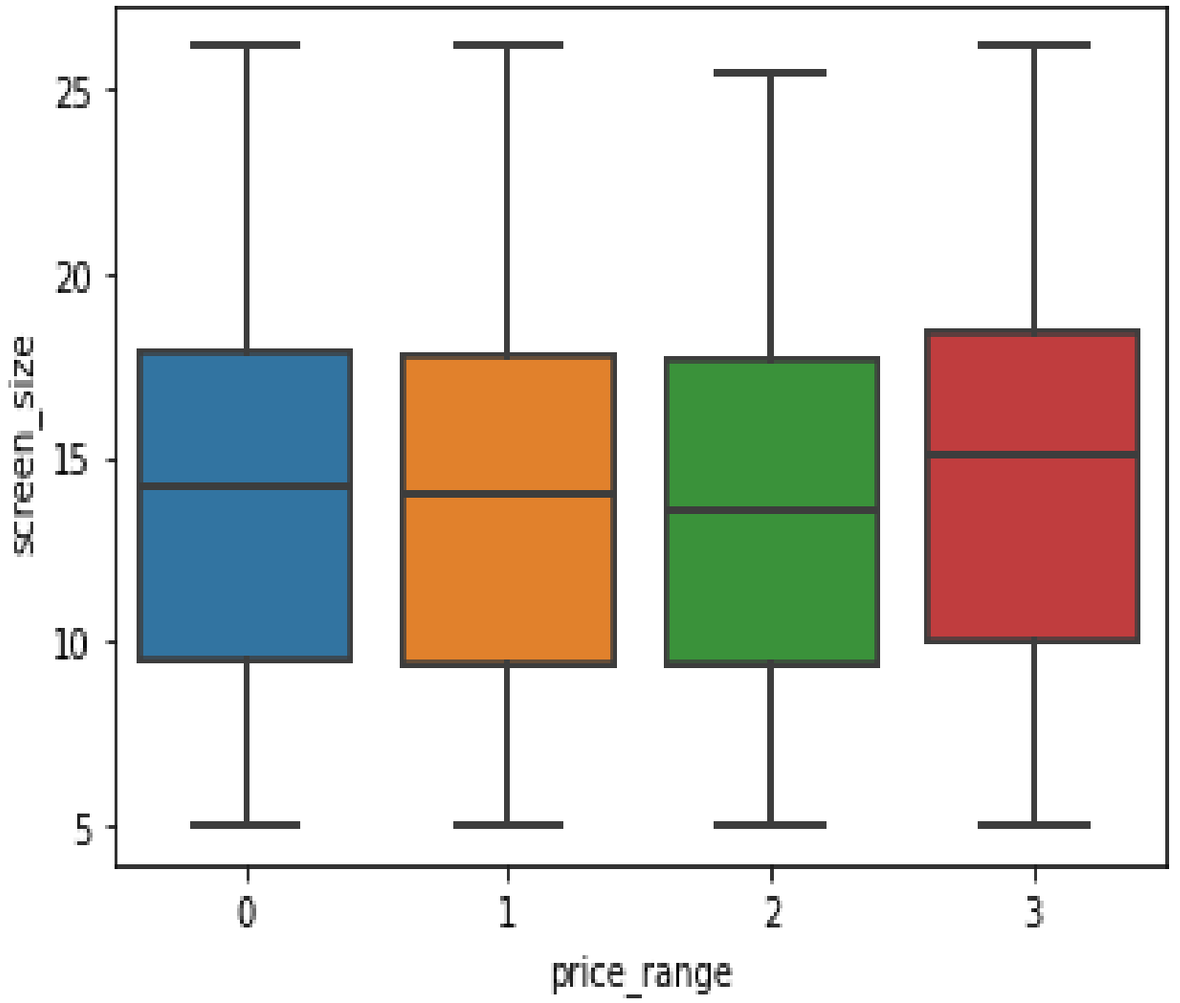



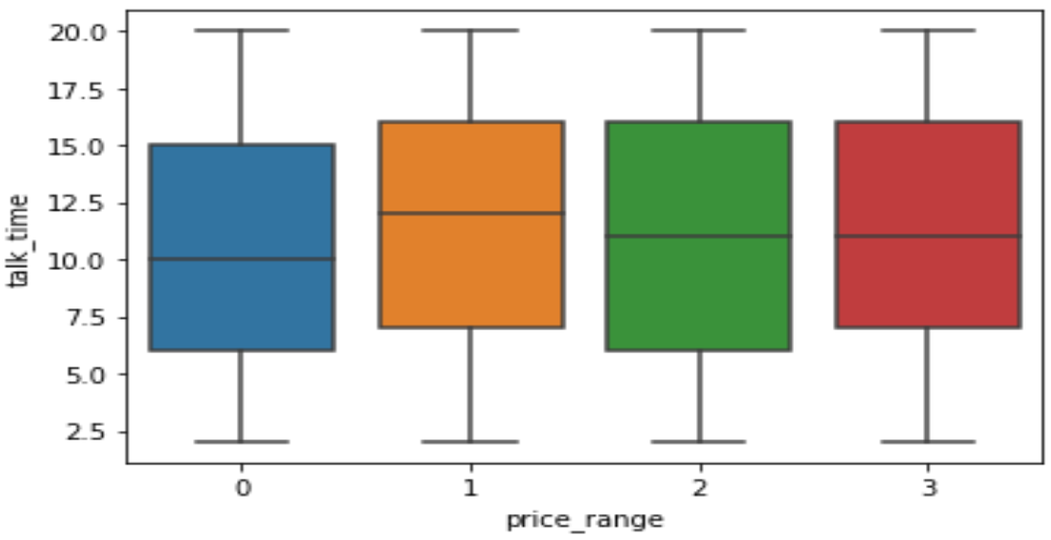

**We can clearly see in the graph with the increase in the ram there is increase in the price range.**

**25 to 50 % mobile depth feature belong to 1 and 3 mobile price range.**

**50% internal memory feature belongs to 1 and 3 mobile price range.**




**50% data of primary megapixel camera feature belong to price-range 1,2,3.**

**50% longest talk time feature that belong to price range 1,2,3.**

**We can see in the graph Screen size feature is max for 3 mobile price range.**

# Machine Learning Modelling

# Model Selection and Evaluation :

Before building a models we performed the train test split. We kept 15% of the data for test and remaining 85% of the data for training the model.

We compared 4 algorithms and evaluated them based on the overall accuracy score and the recall of the individual classes.

- Accuracy is the ratio of the total number of correct predictions and the total number of predictions.
- The recall is the measure of our model correctly identifying True Positives.

1)Logistic regression ML algorithm

2)LightGBM Classifier

3) Decision Tree classifier

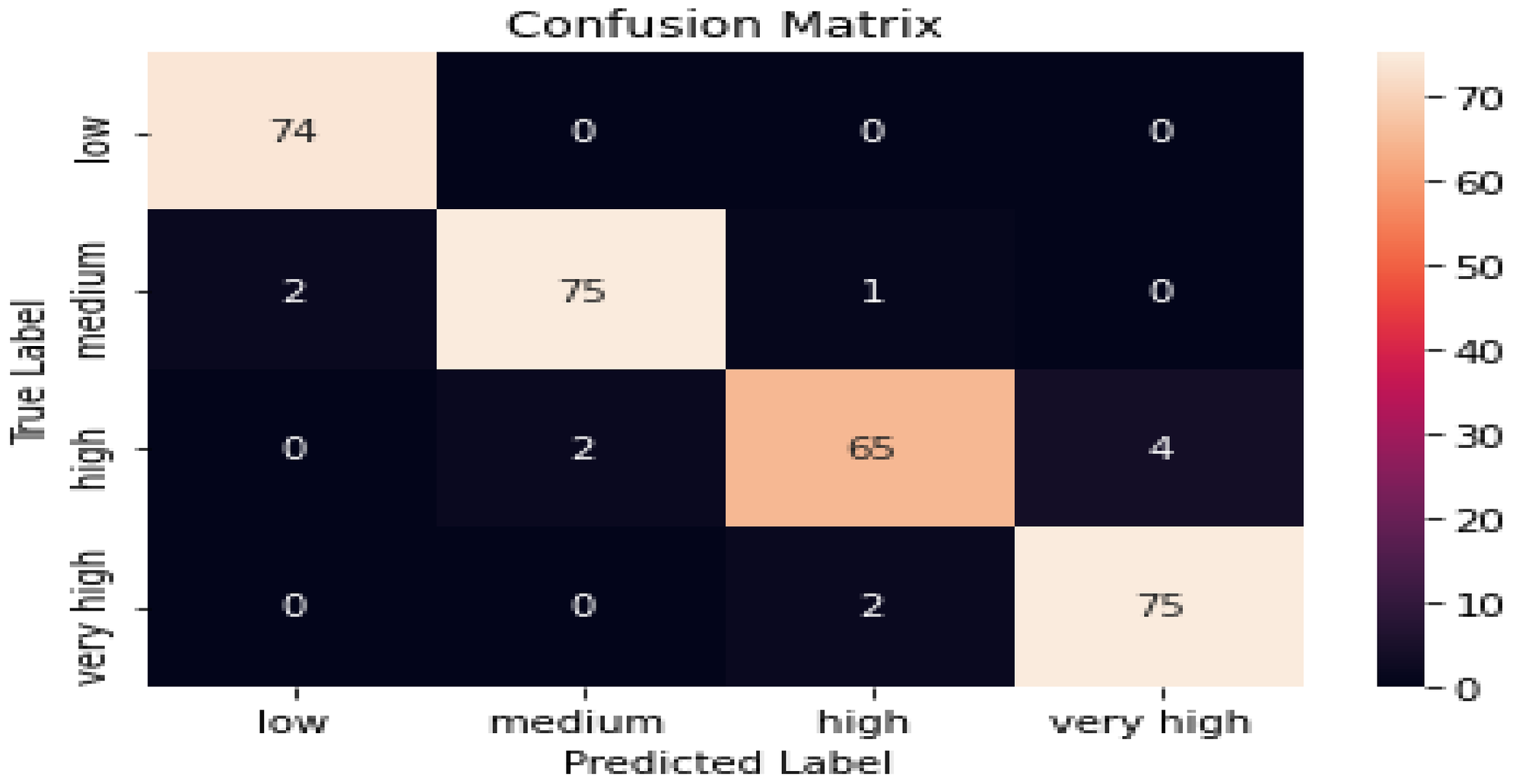
4) Random Forest classifier

# Standardization

**Standardization is an important technique that is mostly performed as a preprocessing step before many Machine Learning models, to standardize the range of features of input data set.**

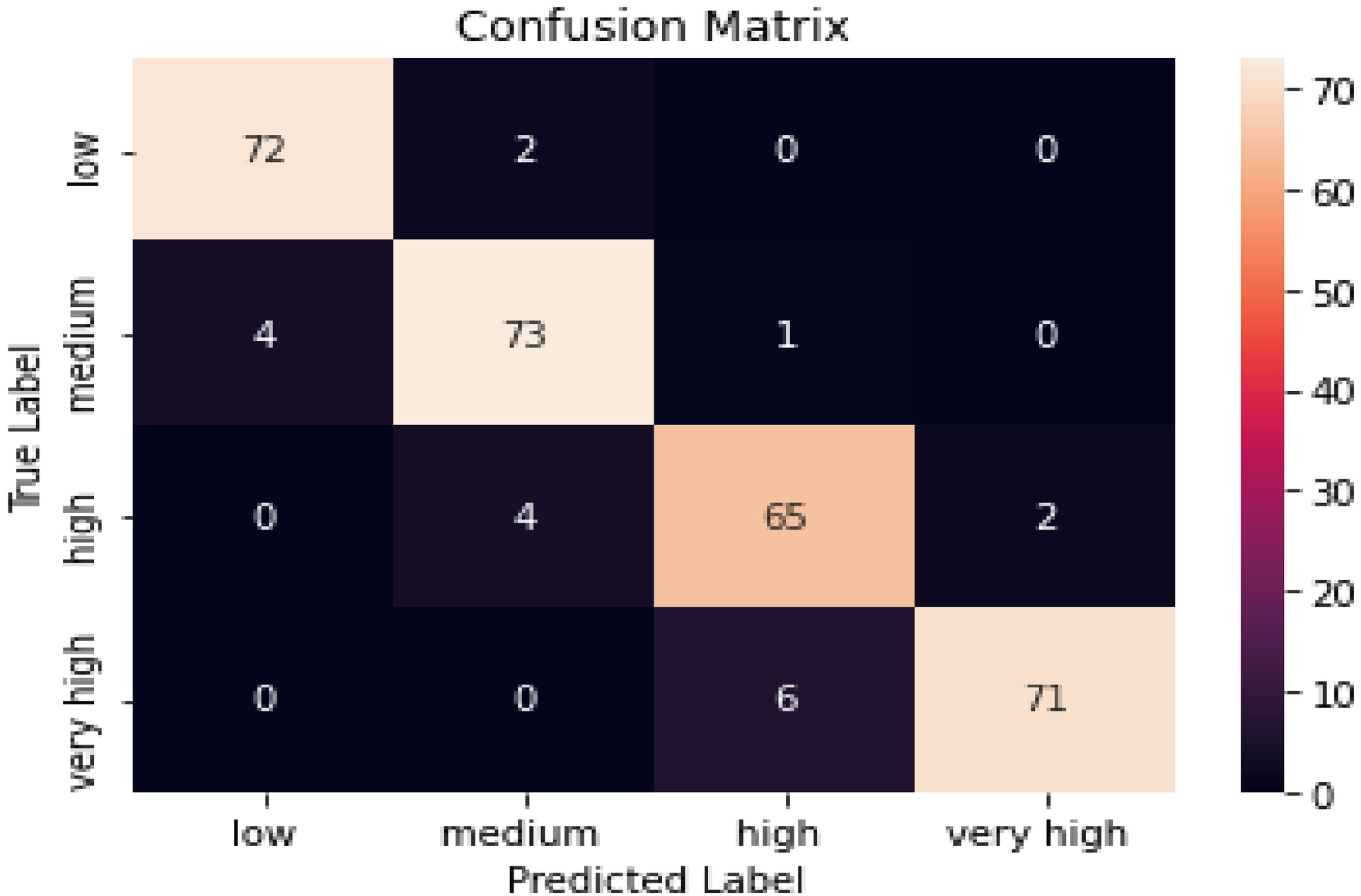
```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()

x_train=sc.fit_transform(x_train)
x_test=sc.fit_transform(x_test)

x_train=pd.DataFrame(x_train,columns=x.columns)
x_test=pd.DataFrame(x_test,columns=x.columns)
```

# Implementing Logistic regression ML algorithm for classification



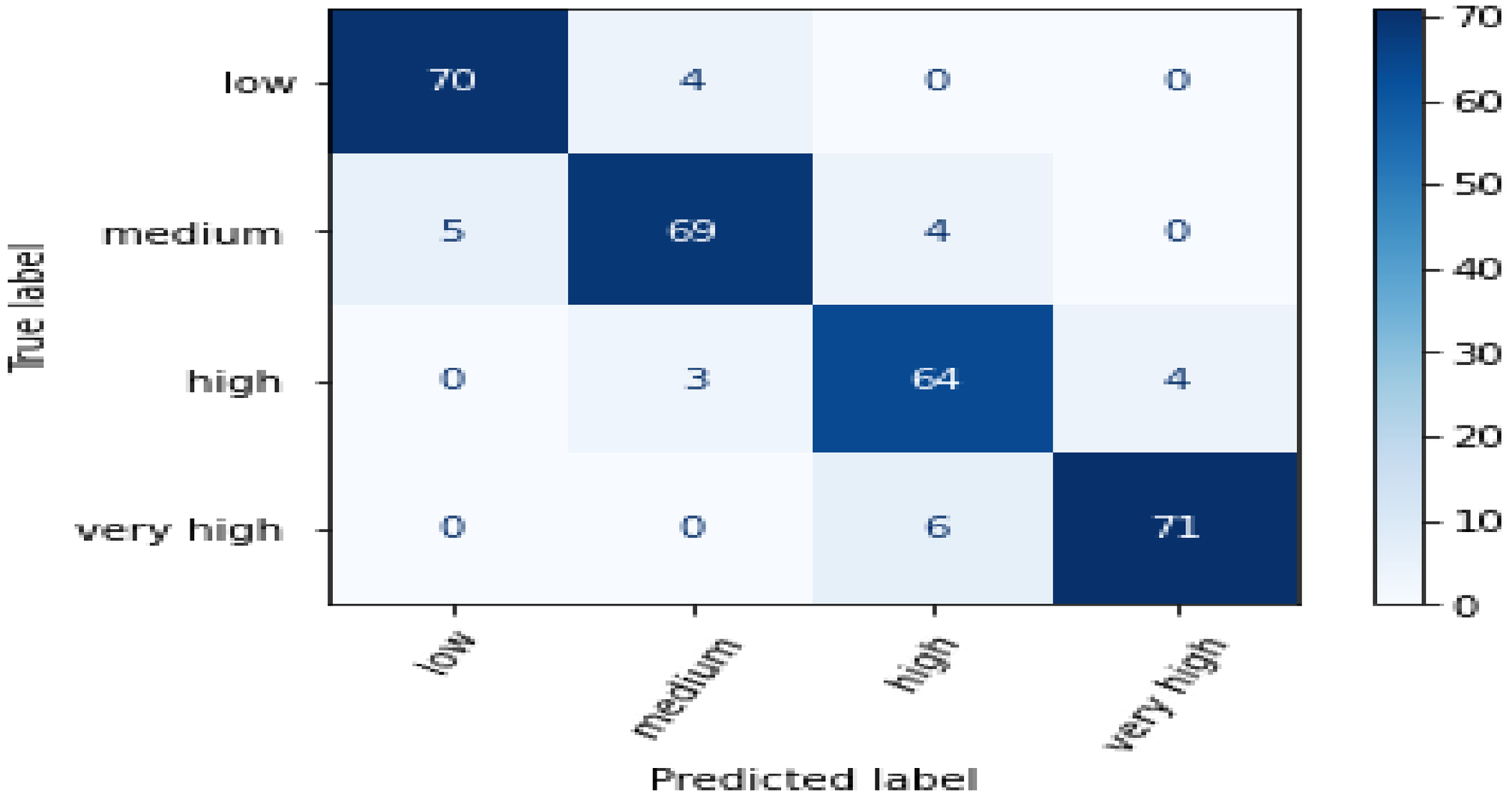

**The Diagonal labels are true predicted labels .All other labels are falsely predicted variables .**

# Implementing LightGBM Classifier model



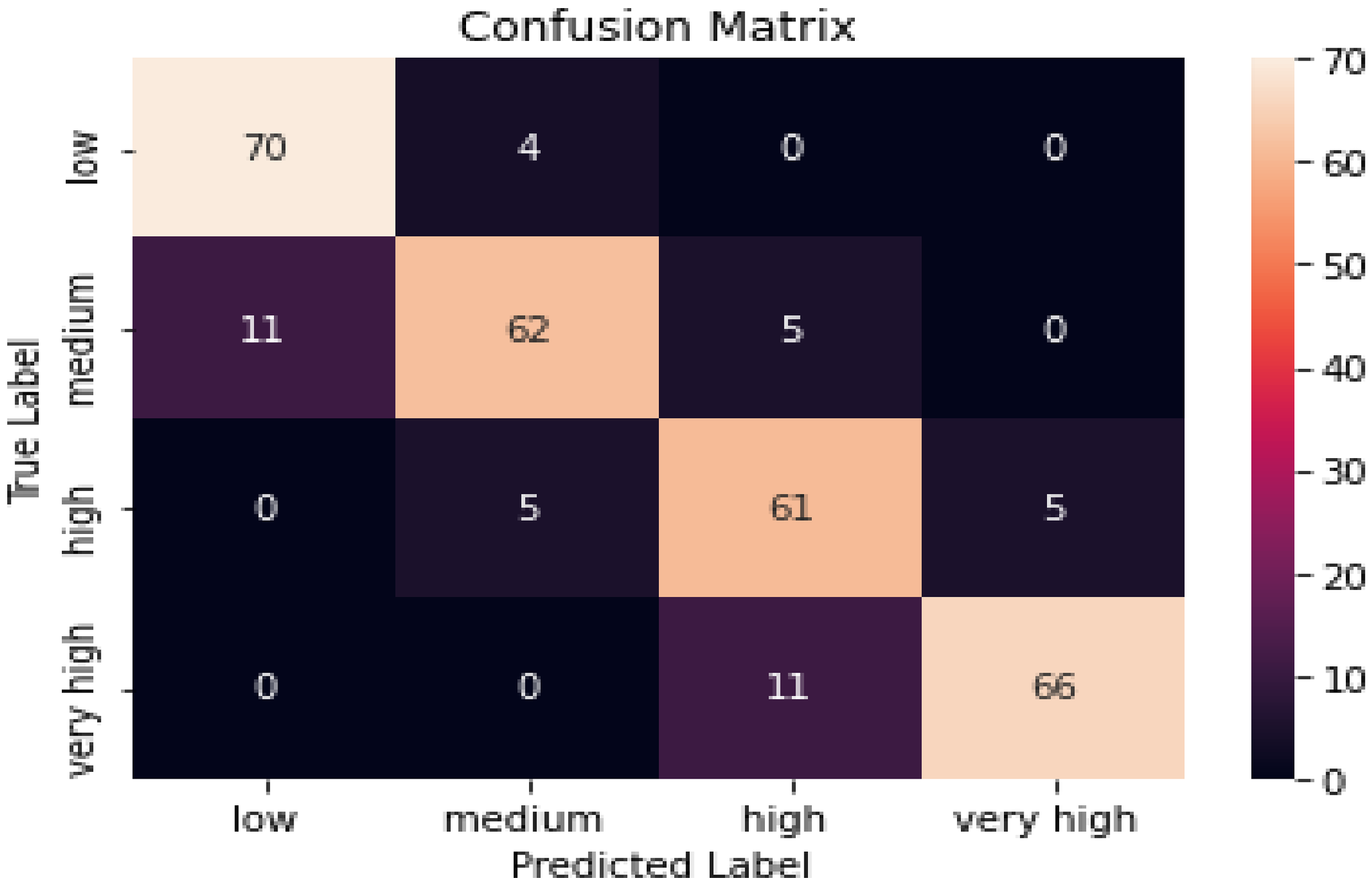
# Implementing Random Forest Classifier model

t




**The Diagonal labels are true predicted labels .All other labels are falsely predicted variables .**

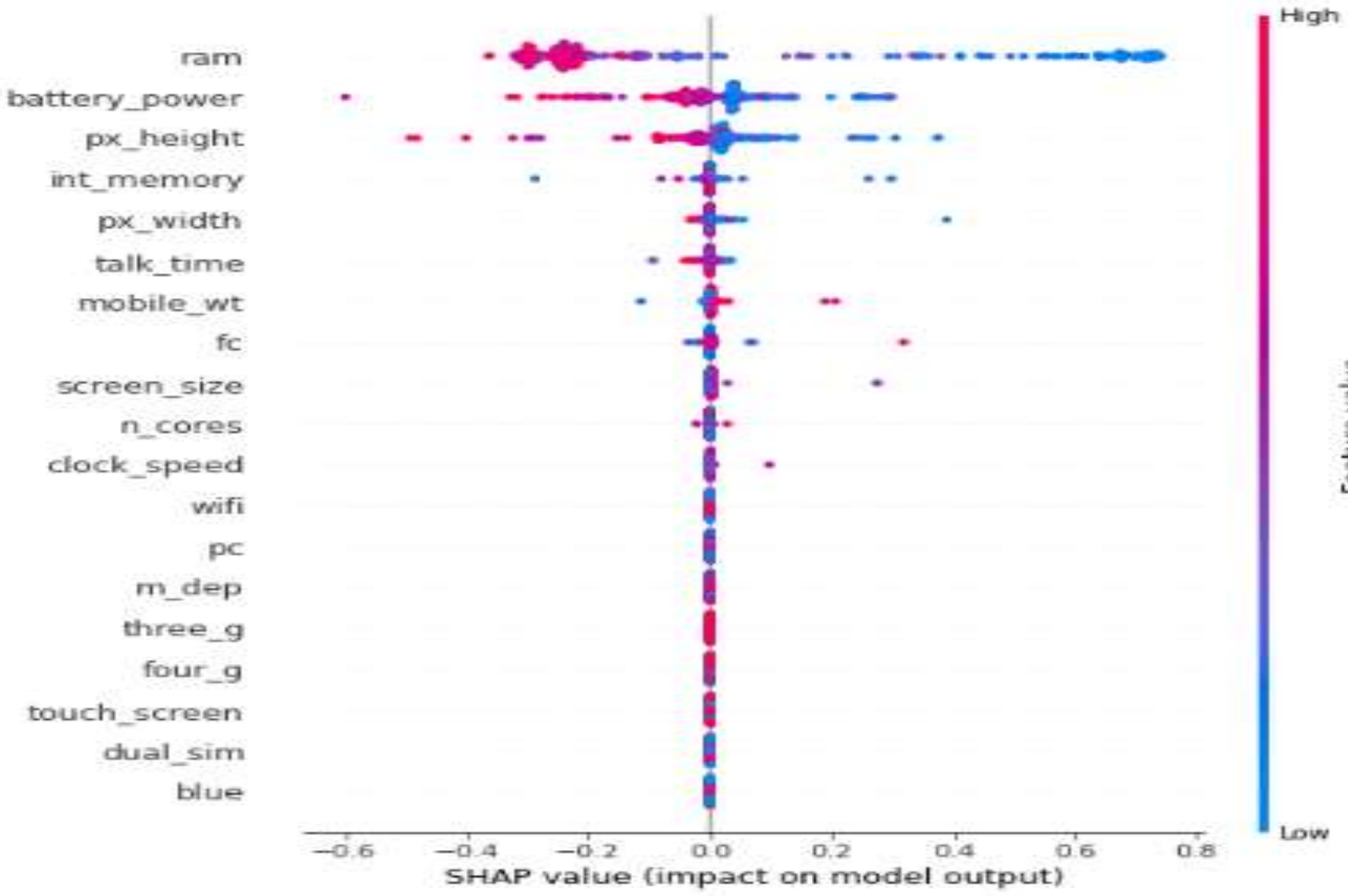
Implementing Decision tree classification ML Algorithm



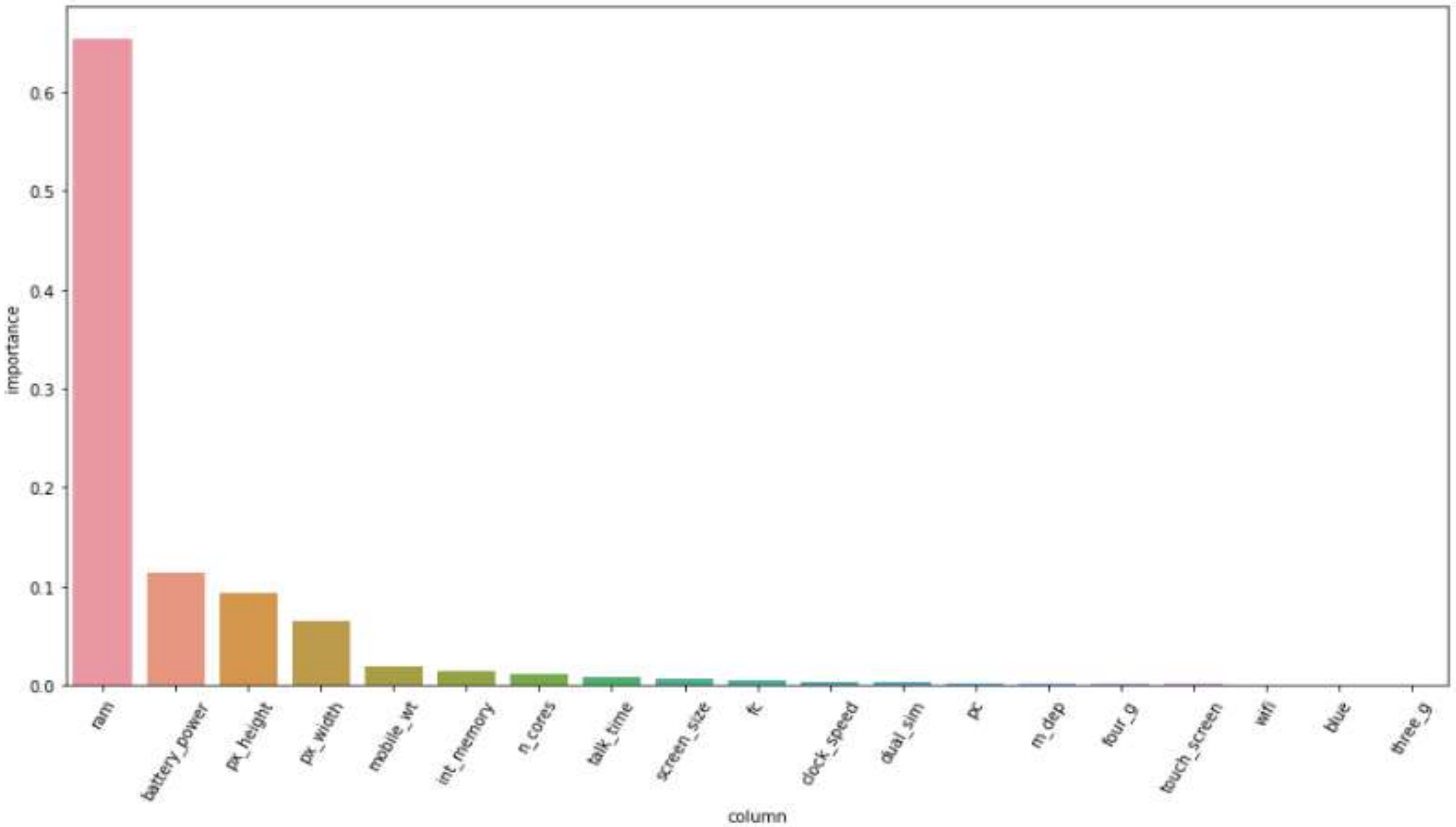
# Evaluation of models:

| Algorithm | Accuracy |
|---|---|
| Logistic regression | 0.9752941176470589 |
| Light gbm | 1.0 |
| Random Forest | 0.8858823529411766 |
| Decision tree | 0.8633333333333333 |

# Interpretation or Justification of features



# feature importance of model



# Conclusion:

1.The given data was cleaned and balanced ,no need to clean data
2.There is not strong relation between any two column ,We perform some feature engineering to reduce column number.
3.Most of the feature like bluetooth,dual_sim, four_g and touch-screen are present in half of the mobiles.
4. Battery power and ram show the most variation along the different price ranges .
5.We splitted the data as to train our model with 85% and test our model with 15% of the total data available.
6.LightGBM and Logistic Regression have highest accuracy compared to other algorithm applied.
7.Ram,Battery_power and mobile_weight are the most important feature when talking about price range.

# Thank You !